Classification of cancer

Field of invention

The present invention relates to a method for classification of cancer in an individual, wherein the microsatellite status and a prognostic marker are determined by examining gene expression patterns. The invention also relates to various methods of treatment of cancer. Additionally, the present invention concerns a pharmaceutical composition for treatment of cancer and uses of the present invention. The invention also relates to an assay for classification of cancer.

Background of invention

Studies of differential gene expression in diseased and normal tissues have been greatly facilitated by the building of large databases of the human genome sequences. Gene expression alterations are important factors in the progression from normal tissue to diseased tissue. In order to obtain a profile of transcriptional status in a certain cell type or tissue, array-based screening of thousands of genes simultaneously is an invaluable tool. Array-based screening even allows for the identification of key genes that alone, or in combination with other genes, regulate the behaviour of a cell or tissue. Candidate genes for future therapeutic intervention may thus also be identified.

Colorectal cancer generally occurs in 1 out of every 20 individuals at some point during their lifetime. In the United States alone about 150,000 new cases are diagnosed each year which amount to 15% of the total number of new cancer diagnoses. Unfortunately, colorectal cancer causes about 56,000 deaths a year in the United States.

The malignant transformation from normal tissue to cancer is believed to be a multistep process. Two molecular pathways are known to be involved in the development of colorectal cancer (Lengauer C, Kinzler KW, Vogelstein B., 1998) namely the microsatellite stable (MSS) pathway and the microsatellite instable (MSI) pathway. MSS is associated with high frequency of allelic losses, abnormalities of cytogenetic nature and abnormal tumor content of DNA. MSI however is associated with defects in the DNA mismatch repair system which leads to increased rate of point mutations and minor chromosomal insertions or deletions.

MSI tumors can be of hereditary or sporadic nature. Ninety percent of MSI tumours are of sporadic origin. Sporadic tumours are presumably MSI due to epigenetic hypermethylation of the MLH1 gene promoter. The hereditary tumours account for 10 % of the MSI tumors. Mutations of for example the MLH1 or MSH 2 genes are often the cause of hereditary tumor development.

The ability of being able to determine the sporadic or hereditary nature of a MSI tumor is highly valuable. In case a tumor is characterized as being MSI , and certain clinical criteria are fulfilled such as age below 50 or three first degree relatives with colon cancer, a screening programme of family members for early diagnosis and treatment of potential colon or endometrial cancer development is initiated. The human and economic costs in relation to screening programmes are severe. Consequently, a need for identifying colon cancers with a hereditary character exists. Further, these patients have a poor prognosis, as they have an increased risk of metachronous colon tumors and a highly increased risk of getting cancer in the endometrium (females), upper urinary tract and a number of other organs. Thus, one may regard the determination of a colon tumor as being sporadic or hereditary as determination of a prognostic factor.

Tumors appearing to be similar – morphologically, histochemically or microscopically – can be profoundly different. They can have different invasive and metastasizing properties, as well as respond differently to therapy. There is thus a need in the art for methods which distinguish tumors and tissues on different bases than are currently in use in the clinic. Determination of microsatellite status using an array-based methodology is faster than conventional DNA based methods, as it does not require microdissection, and forms a set of genes that can be combined with other sets of genes on a colon cancer array that can be used to determine microsatellite status as well as e.g. predict disease course by identifying hereditary cases or other prognostic important factors, and finally predict therapy response.

Summary of invention

In one aspect the present invention relates to a method of classifying cancer in an individual having contracted cancer comprising

in a sample from the individual having contracted cancer determining the microsatellite status of the tumor and

in a sample from the individual having contracted cancer, said sample comprising a plurality of gene expression products the presence and/or amount which forms a pattern, determining from said pattern a prognostic marker, wherein the microsatellite status and the prognostic marker is determined simultaneously or sequentially

classifying said cancer from the microsatellite status and the prognostic marker.

The cancer may be any cancer known to be microsatellite instable in at least a fraction of the cases, such as colon cancer, uterine cancer, ovary cancer, stomach cancer, cancer in the small intestine, cancer in the biliary system, urinary tract cancer, brain cancer or skin cancer. These cancers are part of the spectrum of cancers that belong to the hereditary non-polyposis colon cancer syndrome, but the invention is not limited to this syndrome.

Gene expression patterns may be formed by only a few genes, but it is also a preferred embodiment that a multiplicity of genes form the expression pattern whereby information for classification of cancer can be obtained.

Furthermore, the invention relates to a method for classification of cancer in an individual having contracted cancer, wherein the microsatellite status is determined by a method comprising the steps of

in a sample from the individual having contracted cancer, said sample comprising a plurality of gene expression products the presence and/or amount of which forms a pattern that is indicative of the microsatellite status of said cancer,

determining the presence and/or amount of said gene expression products forming said pattern,

obtaining an indication of the microsatellite status of said cancer in the individual based on the step above.

Yet another aspect of the invention relates to a method for classification cancer in an individual having contracted cancer, wherein the hereditary or sporadic nature is determined by a method comprising the steps of

in a sample from the individual having contracted cancer, said sample comprising a plurality of gene expression products the presence and/or amount of which forms a pattern that is indicative of the hereditary or sporadic nature of said cancer,

determining the presence and/or amount of said gene expression products forming said pattern,

obtaining an indication of the hereditary or sporadic nature of said cancer in the individual based on the step above.

The present invention further concerns a method for treatment of an individual comprising the steps of

selecting an individual having contracted a colon cancer, wherein the microsatellite status is stable, determined according to any of the methods as defined herein

treating the individual with anti cancer drugs .

Another aspect of the present invention relates to a method for treatment of an individual comprising the steps of

selecting an individual having contracted a colon cancer, wherein the microsatellite status is instable, determined according to any of the methods as defined herein

treating the individual with anti cancer drugs.

Yet another aspect of the present invention relates to a method for reducing malignancy of a cell, said method comprising

contacting a tumor cell in question with at least one peptide expressed by at least one gene selected from genes being expressed at least two-fold higher in tumor cells than the amount expressed in said tumor cell in question.

Additionally, the present invention concerns a method for reducing malignancy of a tumor cell in question comprising,

obtaining at least one gene selected from genes being expressed at least two fold lower in tumor cells than the amount expressed in normal cells

introducing said at least one gene into the tumor cell in question in a manner allowing expression of said gene(s).

The invention also relates to a method for reducing malignancy of a cell in question, said method comprising

obtaining at least one nucleotide probe capable of hybridising with at least one gene of a tumor cell in question, said at least one gene being selected from genes being expressed in an amount at least two-fold higher in tumor cells than the amount expressed in normal cells, and

introducing said at least one nucleotide probe into the tumor cell in question in a manner allowing the probe to hybridise to the at least one gene, thereby inhibiting expression of said at least one gene.

In a further aspect the invention relates to a method for producing antibodies against an expression product of a cell from a biological tissue, said method comprising the steps of

obtaining expression product(s) from at least one gene said gene being expressed as defined herein

immunising a mammal with said expression product(s) obtaining antibodies against the expression product.

The present invention also concerns a method for treatment of an individual comprising the steps of

selecting an individual having contracted a colon cancer, wherein the microsatellite status is stable, determined according to any of the methods as defined herein

introducing at least one gene into the tumor cell in a manner allowing expression of said gene(s).

The present invention further relates to a pharmaceutical composition for the treatment of a classified cancer comprising at least one antibody as defined herein.

In yet another aspect the invention concerns a pharmaceutical composition for the treatment of a classified cancer comprising at least one polypeptide as defined herein.

Further, the invention relates to a pharmaceutical composition for the treatment of a classified cancer comprising at least one nucleic acid and/or probe as defined herein.

In an additional aspect the present invention relates to an assay for classification of cancer in an individual having contracted cancer, comprising

at least one marker capable of determining the microsatellite status in a sample and

at least one marker in a sample determining the prognostic marker, wherein the microsatellite status and the prognostic marker is determined simultaneously or sequentially.

Detailed description of the drawings

Figure 1

Unsupervised hierarchical clustering of colorectal tumors based on the 1239 genes with the highest variation across all tumors.

The phylogenetic tree shows the spontaneous clustering of tumor samples and normal biopsies. Germline mutation indicates samples with hereditary mutations in

either *MLH1* or *MSH2* genes. In columns referring to results of immunohistochemistry a plus indicates a positive antibody staining. Tumor location indicates right-sided or left-sided location in the colon of the tumor.

Figure 2

Summary of the performance of the microsatellite instability classifier based on microarray data.

Panel A shows the number of classification errors as a function of the number of genes used. Panel B shows $\log_2$ of the ratio of the distance between a tumor to the centers of the microsatellite instable group and the microsatellite stable tumors. A value of +2 indicates that the distance of a tumor to the microsatellite instable group is 4 times the distance to the microsatellite stable group. Open bars are MSI tumors and solid bars are MSS tumors. Panel C shows the result of the permutation analysis for estimation of the stability of the classifier. This was estimated by generating one hundred new classifiers based on randomly chosen datasets from the 101 tumors each consisting of 30 microsatellite stable and 25 microsatellite instable samples. In each case the classifier was tested with the remaining 46 samples. The performance for each set was evaluated and averaged over all 100 training and test sets.

Figure 3

Classification of MSI tumors as hereditary or sporadic cases based on two genes.

Panel A shows the number of classification errors as a function of the number of genes used. In crossvalidation we found a minimum number of one error using two genes and adding more genes increased the number of errors to a maximum number of twelve. Both genes were used in at least 36 of the 37 crossvalidation loops. Panel B shows $\log_2$ of the ratio of the distance between a tumor to the centers of the sporadic microsatellite instable group and the hereditary microsatellite instable group. Panel C shows microarray signal values for *MLH1* and *PIWIL1* genes for all tumors. Asterisk indicates the misclassified tumor

Figure 4

Classification of microsatellite-instability status based on real-time PCR.

Panel A shows a cluster analysis of 18 of the 101 tumors samples and 9 genes based on the microarray data and compared to real-time PCR data from same samples and genes. Dark colors indicate relative low expression and light/light grey color palette high expression. Panel B shows the result of 47 new independent samples based on PCR data from 7 of the 9 genes. Relative distances are explained in the legend to figure 2. The two misclassified tumors are indicated with an asterisk. For PCR primers and hybridization probes see supplement to methods.

Figure 5

Kaplan-Meier estimates of crude survival among patient with Stage II and Stage III colorectal cancer according to microsatellite status of the tumor, determined by gene expression. Open triangles indicate censored samples. The patients left at risk are denoted in brackets. The P values were calculated with use of the log-rank test.

Figure 6

Phylogenetic tree resulting from unsupervised hierarchical clustering. Clusteranalysis of colon specimens with associated clinicopathological features.

Figure 7

Multidimentional scaling plot showing distances between groups of tumors.

Figure 8

Performance of prediction of survival before and after separation in MSI-H and MSS

Figure 9

Performance of the classifier for identification of hereditary disease.

Figure 10

Kaplan Meier estimates of overall survival among patients with Dukes' B and Dukes' C colon cancer according to microsatellite-instability status of the tumor, determined by gene expression.

Detailed description of the invention

Classification of cancer

The present inventors have, using large-scale array-based screenings, found a pool of genes, the expression products of which may be used to classify cancer in an individual. The presence of expression products and level of expression products provides an expression pattern which is correlated to a specific status and/or prognostic marker of the cancer. Characterization of the genes or functional analysis of the gene expression products as such is not required to classify the cancer based on the present method. Thus, the expression products of the plurality of genes can be used as markers for the classification of disease.

One aspect of the present invention concerns a method for classifying cancer in an individual having contracted cancer by determining the microsatellite status and a prognostic marker in a sample. Determination of the microsatellite status and the prognostic marker may be performed simultaneously or sequentially. In one embodiment of the present invention the microsatellite status is determined. The prognostic marker is determined in a sample, wherein the presence and/or the amount of a number of gene expression products form a pattern wherefrom the prognostic marker is determined. Based on the information gathered from the microsatellite status and the prognostic marker the cancer can be classified. In a preferred embodiment the prognostic marker is the hereditary or sporadic nature of the cancer. The hereditary or sporadic nature of the cancer can be determined through a number of steps comprising determining the presence and/or amount of gene expression products forming a pattern in a sample. The sample comprises a number of gene expression products the presence and/or amount of which forms a pattern that is indicative of the hereditary or sporadic nature of the cancer. Hereby, an indication of the hereditary or sporadic nature of the cancer is obtained.

In one embodiment of the invention the microsatellite status is determined using conventional analysis of microsatellite status as described elsewhere herein.

In another embodiment of the present invention the microsatellite status is determined by gene expression patterns wherein the presence and/or the amount of the gene expression products form a pattern that is indicative of the microsatellite status.

Classification of cancer provides knowledge of the survival chances of an individual having contracted cancer. In case of cancer which according to the present invention has been classified as a hereditary cancer, screening programmes of family members to the individual having the classified cancer can be initiated. Such screening programmes can comprise conventional screening programmes employing sequencing and other methods as described elsewhere. Thus, individuals at risk of developing cancer may be identified and action taken accordingly to detect developing cancer at an early stage of the disease greatly improving the chances of successful intervention and thus survival rates.

Classification of cancer also provides insights on which sort of treatment should be offered to the individual having contracted cancer, thus providing an improved treatment response of the individual. Likewise, the individual may be spared treatment that is inefficient in treating the particular class of cancer and thus spare the individual severe side effects associated with treatment that may even not be suitable for the class of cancer.

Microsatellite status

The use of highly variable repetitive sequences found in microsatellite regions adjacent to genes or other areas of interest may be used as markers for linkage analysis, DNA fingerprinting, or other diagnostic application.

Microsatellites are defined as loci (or regions within DNA sequences) where short sequences of DNA are repeated in tandem repeats. This means that the sequences are repeated one right after the other. The lengths of sequences used most often are di-, tri-, or tetra-nucleotides. At the same location within the genomic DNA the number of times the sequence (ex. AC) is repeated often varies between individuals, within populations, and/or between species. Due to the many repeats the microsatellites are prone to alter if there is a reduced repair of mismatches in the genome.In the present invention the traditional method of determining microsatellite status by employing microsatellite markers is replaced by determination of gene expression patterns.

An important factor in multi-step carcinogenesis is genomic instability. The development of some cancer forms is known to follow two distinct molecular routes. One

route is the microsatellite stable, MSS, (and chromosomal instable pathway) which is often associated with a high frequency of allelic losses, cytogenetic abnormalities and abnormal DNA tumor contents. The second route is the microsatellite instable pathway MSI that is characterized by defects in the DNA mismatch repair system which leads to a high rate of point mutations and small chromosomal insertions and deletions. The small chromosomal insertions and deletions can be detected as mono and dinucleotide repeats (Boland CR, Thibodeau SN, Hamilton SR, et al., Cancer Res 1998;58(22):5248-57).

One aspect of the present invention relates to the classification of cancer in an individual having contracted cancer by determining the microsatellite status and a prognostic marker. One embodiment of the invention relates to microsatellite status determined by conventional methods employing microsatellite analysis as described above. Another embodiment of the invention relates to establishing the microsatellite status by determining the presence and/or amount of gene expression products of a sample which comprises a plurality of gene expression products forming a pattern which is indicative of the microsatellite status.

The expression products of genes according to the present invention are not necessarily identical to the genes that are analysed by microsatellite markers in conventional methods of determining microsatellite status. The pattern of the gene expression products according to the present invention however correlates with information on microsatellite status that can be obtained using traditional methods.

The determination of the microsatellite status and the prognostic marker of the cancer may be performed sequentially. However, the determinations may also be performed simultaneously.

Prognostic marker

Together with knowledge of the microsatellite status in a sample of an individual having contracted cancer a prognostic marker is employed for classifying the cancer. The prognostic marker may be any marker that provides knowledge of the cancer type when combined with knowledge of microsatellite status. Consequently the prognostic marker may provide additional information on the cancer type when the microsatellite status is stable and similarly when the microsatellite status is

instable. In a preferred embodiment of the present invention the prognostic marker is the hereditary or sporadic nature of a cancer given that the microsatellite status is instable. The prognostic marker may in another embodiment be a prognostic marker for any feature or trait that provides further possibilities of classifying cancer.

The prognostic marker is determined in a sample comprising a number of gene expression products wherein the presence and/or amounts of gene expression products form a pattern that is indicative of the prognostic marker.

Hereditary and sporadic nature of cancer

Hereditary nonpolyposis colon cancer (HNPCC) is a hereditary cancer syndrome which carries a very high risk of colon cancer and an above-normal risk of other cancers (uterus, ovary, stomach, small intestine, biliary system, urinary tract, brain, and skin). The HNPCC syndrome is due to mutation in a gene in the DNA mismatch repair system, usually the MLH1 or MSH2 gene or less often the MSH6 or PMS2 genes. Families with HNPCC account for about 5% of all cases of colon cancer and typically have the following features (called the Amsterdam clinical criteria):

Three or more first relative family members with colorectal cancer; affected family members in two or more generations; and at least one person with colon cancer diagnosed before the age of 50.

The highest risk with HNPCC is for colon cancer. A person with HNPCC has about an 80% lifetime risk of colon cancer. Two-thirds of these tumors occur in the proximal colon. Women with HNPCC have a 20-60% lifetime risk of endometrial cancer. In HNPCC, the gastric cancer is usually intestinal-type adenocarcinoma. The ovarian cancer in HNPCC may be diagnosed before age 40. Other HNPCC-related cancers have characteristic features: the urinary tract cancers are transitional carcinoma of the ureter and renal pelvis; the small bowel cancer is most common in the duodenum and jejunum; and the most common type of brain tumor is glioblastoma.

The diagnosis of HNPCC may be made on the basis of the Amsterdam clinical criteria (listed above) or on the basis of molecular genetic testing for mutations in a mismatch repair gene (MLH1, MSH2, MSH6 or PMS2). Mutations in MLH1 and MSH2 account for 90% of HNPCC. Mutations in MSH6 and PMS2 account for the rest.

HNPCC is inherited in an autosomal dominant manner. Each child of an individual with HNPCC has a 50% chance of inheriting the mutation. Most people diagnosed with HNPCC have inherited the condition from a parent. However, not all individuals with an HNPCC gene mutation have a parent who had cancer. Prenatal diagnosis for pregnancies at increased risk for HNPCC is possible.

In tumors that are microsatellite instable it is often found that the DNA mismatch repair proteins that are encoded by the *MLH1* or *MSH2* genes are inactivated. In case of microsatellite instable hereditary non-polyposis colorectal cancers germline mutation in MLH1 and MSH2 and somatic loss of function of the normal allele has been found to be associated with the disease.

For most sporadic MSI tumors epigenetic hypermethylation of the MLH1 promoter can be found to be associated with the cancer (Cunningham JM, Christensen ER, Tester DJ, et al., Cancer Res 1998;58(15):3455-60., Kane MF, Loda M, Gaida GM, et al., Cancer Res 1997;57(5):808-11., Herman JG, Umar A, Polyak K, et al., Proc Natl Acad Sci U S A 1998;95(12):6870-5., Kuismanen SA, Holmberg MT, Salovaara R, de la Chapelle A, Peltomaki P., Am J Pathol 2000;156(5):1773-9).

Forms of cancer

Cancer leads to a change in the expression of one or more genes. The methods according to the invention may be used for classifying cancer according to the microsatellite status and/or the hereditary or sporadic nature of the cancer. Thus, the cancer may be any malignant condition in which genomic instability is involved in the development of cancer, such as cancers related to hereditary non-polyposis colorectal cancer, such as endometrial cancer, gastric cancer, small bowel cancer, ovarian cancer, kidney cancer, pelvic renal cancer or tumors of the nervous system, such as glioblastoma.

One particular form of cancer according to the present invention is that of the colon/rectum.

The cancer may be of any tumor type, such as an adenocarcinoma, a carcinoma, a teratoma, a sarcoma, and/or a lymphoma.

In relation to the gastro-intestinal tract, the biological condition may also be colitis ulcerosa, Mb. Crohn, diverticulitis, adenomas.

Colorectal tumors

The data presented herein relates to colorectal tumors and therefore the description has focused on the gene expression level as one manner of identifying genes involved in the prediction of survival in cancer tissue. The malignant progression of cancer of colon or rectum may be described using Dukes stages where normal mucosa may progress to Dukes A superficial tumors to Dukes B, slightly invasive tumors, to Dukes C that have spread to lymphnodes and finally to Dukes D that have metastasized to other organs.

The grade of a tumor can also be expressed on a scale of I-IV. The grade reflects the cytological appearance of the cells. Grade I cells are almost normal, whereas grade II cells deviate slightly from normal. Grade III appear clearly abnormal, whereas grade IV cells are highly abnormal.

The phrase colon cancer is in this application meant to be equivalent to the phrase colorectal cancer. Colon cancers may be located in the right side of the colon, the left side of the colon, the transverse part of the colon and/or in the rectum.

Samples

The samples according to the present invention may be any cancer tissue.

The sample may be in a form suitable to allow analysis by the skilled artisan, such as a biopsy of the tissue, or a superficial sample scraped from the tissue. In one embodiment of the invention it is preferred that the sample is from a resected colon cancer tumor. In another embodiment the sample may be prepared by forming a suspension of cells made from the tissue. The sample may, however, also be an extract obtained from the tissue or obtained from a cell suspension made from the tissue. The sample may be fresh or frozen, or treated with chemicals.

Expression pattern

Expression of one gene or more genes in a sample forms a pattern that is characteristic of the state of the cell. In a sample from an individual having contracted cancer a plurality of gene expression products are present. By expression pattern is meant

the presence of a combination of a number of expression products and/or the amount of expression products specific for a given biological condition, such as cancer. The pattern is produced by determining the expression products of selected genes that together reveals a pattern that is indicative of the biological condition. Thus, a selection of the genes that carry information about a specific condition is developed. Selection of the genes is achieved by analyzing large numbers of genes and their expression products to find the genes that will enable the desired differentiation between various conditions, such as microsatellite status (MSS or MSI) and/or prognostic marker, such as for example the sporadic or hereditary nature of a given cancer sample. The criteria for selection of the best genes for the pattern to be indicative of given biological conditions include confidence levels i.e. how accurate are the selected genes forming an expression pattern in giving correct information of the biological condition. Thus, in one aspect of the present invention a specific pattern of gene expression profiles can be used to determine the microsatellite status in the sample. In a second aspect of the present invention the microsatellite status is determined and a specific pattern of the presence of a plurality of gene expression products and/or amount wherefrom a prognostic marker is determined.

Determination of the microsatellite status employing gene expression patterns

One aspect of the invention specifically relates to a method for determining the microsatellite status in a sample of an individual having contracted cancer based on determination of the expression pattern of at least two genes, such as at least three genes, such as at least four genes, such as at least 5 genes, such as at least 6 genes, such as at least 7 genes, such as at least 8 genes, such as at least 9 genes, such as at least 10 genes, such as at least 15 genes, such as at least 20 genes, such as at least 30 genes, such as at least 40 genes, such as at least 50 genes, such as at least 60 genes, such as at least 70 genes, such as at least 80 genes, such as at least 90 genes, such as at least 126 genes selected from the group of genes listed in Table 1 below

Table 1

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
chemokine (C-C motif) ligand 5	NM_002985	CCL5	1
tryptophanyl-tRNA synthetase	NM_004184	WARS	2
proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)	NM_006263	PSME1	3
bone marrow stromal cell antigen 2	NM_004335	BST2	4
ubiquitin-conjugating enzyme E2L 6	NM_004223	UBE2L6	5

A kinase (PRKA) anchor protein 1	NM_003488	AKAP1	6
proteasome (prosome, macropain) activator subunit 2 (PA28 beta)	NM_002818	PSME2	7
carcinoembryonic antigen-related cell adhesion molecule 5	NM_004363	CEACAM5	8
FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)	NM_005766	FARP1	9
myosin X	NM_012334	MYO10	10
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
autocrine motility factor receptor	NM_001144	AMFR	12
dimethylarginine dimethylaminohydrolase 2	NM_013974	DDAH2	13
tumor necrosis factor, alpha-induced protein 2	NM_006291	TNFAIP2	14
mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	NM_000249	MLH1	15
thymidylate synthetase	NM_001071	TYMS	16
intercellular adhesion molecule 1 (CD54), human rhinovirus receptor	NM_000201	ICAM1	17
general transcription factor IIA, 2, 12kDa	NM_004492	GTF2A2	18
Rho-associated, coiled-coil containing protein kinase 2	NM_004850	ROCK2	19
ATP binding protein associated with cell differentiation	NM_005783	TXNDC9	20
NCK adaptor protein 2	NM_003581	NCK2	21
phytanoyl-CoA hydroxylase (Refsum disease)	NM_006214	PHYH	22
metastais-associated gene family, member 2	NM_004739	MTA2	23
amiloride binding protein 1 (amine oxidase (copper-containing))	NM_001091	ABP1	24
biliverdin reductase A	NM_000712	BLVRA	25
phospholipase C, beta 4	NM_000933	PLCB4	26
chemokine (C-X-C motif) ligand 9	NM_002416	CXCL9	27
purine-rich element binding protein A	NM_005859	PURA	28
quinolinate phosphoribosyltransferase (nicotinate-nucleotide pyrophosphorylase (carboxylating))	NM_014298	QPRT	29
retinoic acid receptor responder (tazarotene induced) 3	NM_004585	RARRES3	30
chemokine (C-C motif) ligand 4	NM_002984	CCL4	31
forkhead box O3A	NM_001455	FOXO3A	32
interferon, alpha-inducible protein (clone IFI-6-16)	NM_002038	G1P3	34
	NM_022873		123
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
metallothionein 1G	NM_005950	MT1G	36
	NM_005950		
tumor necrosis factor receptor superfamily, member 6	NM_000043	TNFRSF6	37
	NM_152877		133
	NM_152876		132
	NM_152875		134
	NM_152872		130
	NM_152873		33
	NM_152871		129
	NM_152874		131
endothelial cell growth factor 1 (platelet-derived)	NM_001953	ECGF1	38
SCO cytochrome oxidase deficient homolog 2 (yeast)	NM_005138	SCO2	39
chemokine (C-X-C motif) ligand 13 (B-cell chemoattractant)	NM_006419	CXCL13	40

Granulysin	NM_006433	GNLY	41
CD2 antigen (p50), sheep red blood cell receptor	NM_001767	CD2	42
splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43
teratocarcinoma-derived growth factor 1	NM_003212	TDGF1	44
metallothionein 1H	NM_005951	MT1H	45
cytochrome P450, family 2, subfamily B, polypeptide 6	NM_000767	CYP2B6	46
tumor necrosis factor (ligand) superfamily, member 9	NM_003811	TNFSF9	47
RNA binding motif protein 12	NM_006047 NM_006047	RBM12	48
heat shock 105kDa/110kDa protein 1	NM_006644	HSPH1	49
staufen, RNA binding protein (Drosophila)	NM_004602 NM_017452 NM_017453	STAU	50 125 126
lymphocyte antigen 6 complex, locus G6D	NM_021246	LY6G6D	51
calcium binding protein P22	NM_007236	CHP	52
CDC14 cell division cycle 14 homolog B (S. cerevisiae)	NM_003671 NM_033331	CDC14B	53 115
epiplakin 1	XM_372063	EPPK1	54
metallothionein 1X	NM_005952	MT1X	55
transforming growth factor, beta receptor II (70/80kDa)	NM_003242	TGFBR2	56
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 124
transmembrane 4 superfamily member 6	NM_003270	TM4SF6	58
pleckstrin homology domain containing, family B (evectins) member 1	NM_021200	PLEKHB1	59
apolipoprotein L, 1	NM_003661 NM_145343	APOL1	60 120
indoleamine-pyrrole 2,3 dioxygenase	NM_002164	INDO	61
forkhead box A2	NM_021784	FOXA2	62
granzyme H (cathepsin G-like 2, protein h-CCPX)	NM_033423	GZMH	63
baculoviral IAP repeat-containing 3	NM_001165	BIRC3	64
Homo sapiens metallothionein 1H-like protein		AF333388 (Hs 382039)	135
KIAA0182 protein	NM_014615	KIAA0182	117
G protein-coupled receptor 56	NM_005682	GPR56	65

	NM_201524		116
metallothionein 2A	NM_005953	MT2A	66
F-box only protein 21	NM_015002	FBXO21	67
erythrocyte membrane protein band 4.1-like 1	NM_012156, NM_012156	EPB41L1	68
hypothetical protein MGC21416	NM_173834	MGC21416	69
protein O-fucosyltransferase 1	NM_015352, NM_015352	POFUT1	70
metallothionein 1E (functional)	NM_175617	MT1E	71
troponin T1, skeletal, slow	NM_003283	TNNT1	72
chimerin (chimaerin) 2	NM_004067	CHN2	73
heterogeneous nuclear ribonucleoprotein H1 (H)	NM_005520	HNRPH1	74
ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle	NM_004046	ATP5A1	75
eukaryotic translation initiation factor 5A	NM_001970	EIF5A	76
perforin 1 (pore forming protein)	NM_005041	PRF1	77
OGT(O-Glc-NAc transferase)-interacting protein 106 KDa	NM_014965	OIP106	78
DEAD (Asp-Glu-Ala-Asp) box polypeptide 27	NM_017895	DDX27	79
vacuolar protein sorting 35 (yeast)	NM_018206	VPS35	80
tripartite motif-containing 44	NM_017583	TRIM44	81
transmembrane, prostate androgen induced RNA	NM_020182	TMEPAI	82
	NM_199169		127
	NM_199170		128
dynein, cytoplasmic, light polypeptide 2A	NM_014183	DNCL2A	83
	NM_177953		122
leucine aminopeptidase 3	NM_015907	LAP3	84
chromosome 20 open reading frame 35	NM_018478	C20orf35	85
	NM_033542		118
solute carrier family 38, member 1	NM_030674	SLC38A1	86
CGI-85 protein	NM_016028	CGI-85	87
death associated transcription factor 1	NM_022105,	DATF1	88
	NM_080796		121
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
sestrin 1	NM_014454	SESN1	90
hypothetical protein FLJ20315	NM_017763	FLJ20315	91
hypothetical protein FLJ20647	NM_017918	FLJ20647	92
membrane protein expressed in epithelial-like lung adenocarcinoma	NM_024792	CT120	93
DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide	NM_014314	RIG-I	94
keratin 23 (histone deacetylase inducible)	NM_015515,	KRT23	95
UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-	NM_007210	GALNT6	96

acetylgalactosaminyltransferase 6 (GalNAc-T6)			
aryl hydrocarbon receptor nuclear translocator-like 2	NM_020183	ARNTL2	97
apobec-1 complementation factor	NM_014576, NM_138932	ACF	98 119
hypothetical protein FLJ20232	NM_019008	FLJ20232	99
apolipoprotein L, 2	NM_030882, NM_145343	APOL2	100 120
mitochondrial solute carrier protein	NM_016612	MSCP	101
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

One embodiment of the invention concerning the determination of microsatellite status is based on the expression pattern of at least 2 genes, such as at least 3 genes, such as at least 4 genes, such as at least 5 genes, such as at least 6 genes, such as at least 7 genes, such as at least 8 genes, such as at least 9 genes, such as at least 10 genes, such as at least 15 genes, such as at least 20 genes, such as at least 25 genes selected from the group of genes listed in Table 2.

Table 2

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
chemokine (C-C motif) ligand 5	NM_002985	CCL5	1
tryptophanyl-tRNA synthetase	NM_004184	WARS	2
proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)	NM_006263	PSME1	3
bone marrow stromal cell antigen 2	NM_004335	BST2	4
ubiquitin-conjugating enzyme E2L 6	NM_004223	UBE2L6	5
A kinase (PRKA) anchor protein 1	NM_003488	AKAP1	6
proteasome (prosome, macropain) activator subunit 2 (PA28 beta)	NM_002818	PSME2	7
carcinoembryonic antigen-related cell adhesion molecule 5	NM_004363	CEACAM5	8
FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)	NM_005766	FARP1	9
myosin X	NM_012334	MYO10	10
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
autocrine motility factor receptor	NM_001144	AMFR	12
dimethylarginine dimethylaminohydrolase 2	NM_013974	DDAH2	13
tumor necrosis factor, alpha-induced protein 2	NM_006291	TNFAIP2	14
mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	NM_000249	MLH1	15
thymidylate synthetase	NM_001071	TYMS	16

intercellular adhesion molecule 1 (CD54), human rhinovirus receptor	NM_000201	ICAM1	17
general transcription factor IIA, 2, 12kDa	NM_004492	GTF2A2	18
Rho-associated, coiled-coil containing protein kinase 2	NM_004850	ROCK2	19
ATP binding protein associated with cell differentiation	NM_005783	APACD	20
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 124
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

or from

Table 3

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
NCK adaptor protein 2	NM_003581	NCK2	21
phytanoyl-CoA hydroxylase (Refsum disease)	NM_006214	PHYH	22
metastais-associated gene family, member 2	NM_004739	MTA2	23
amiloride binding protein 1 (amine oxidase (copper-containing))	NM_001091	ABP1	24
biliverdin reductase A	NM_000712	BLVRA	25
phospholipase C, beta 4	NM_000933	PLCB4	26
chemokine (C-X-C motif) ligand 9	NM_002416	CXCL9	27
purine-rich element binding protein A	NM_005859	PURA	28
quinolinate phosphoribosyltransferase (nicotinate-nucleotide pyrophosphorylase (carboxylating))	NM_014298	QPRT	29
retinoic acid receptor responder (tazarotene induced) 3	NM_004585	RARRES3	30
chemokine (C-C motif) ligand 4	NM_002984	CCL4	31
forkhead box O3A	NM_001455	FOXO3A	32
metallothionein 1X	NM_005952	MT1X	55
interferon, alpha-inducible protein (clone IFI-6-16)	NM_002038 NM_022873	G1P3	34 123
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
metallothionein 1G	NM_005950, NM_005950	MT1G	36

	NM_000043	TNFRSF6	37
tumor necrosis factor receptor superfamily, member 6	NM_152877		133
	NM_152876		132
	NM_152875		134
	NM_152872		130
	NM_152873		33
	NM_152871		129
	NM_152874		131
endothelial cell growth factor 1 (platelet-derived)	NM_001953	ECGF1	38
SCO cytochrome oxidase deficient homolog 2 (yeast)	NM_005138	SCO2	39
chemokine (C-X-C motif) ligand 13 (B-cell chemoattractant)	NM_006419	CXCL13	40
Granulysin	NM_006433	GNLY	41
splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43
protein kinase C binding protein 1	NM_012408	PRKCBP1	57
	NM_183047		124
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

or from

Table 4

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
CD2 antigen (p50), sheep red blood cell receptor	NM_001767	CD2	42
splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43
teratocarcinoma-derived growth factor 1	NM_003212	TDGF1	44
metallothionein 1H	NM_005951	MT1H	45
cytochrome P450, family 2, subfamily B, polypeptide 6	NM_000767	CYP2B6	46
tumor necrosis factor (ligand) superfamily, member 9	NM_003811	TNFSF9	47
RNA binding motif protein 12	NM_006047, NM_006047	RBM12	48
heat shock 105kDa/110kDa protein 1	NM_006644	HSPH1	49
staufen, RNA binding protein (Drosophila)	NM_004602	STAU	50
	NM_017452		125
	NM_017453		126
lymphocyte antigen 6 complex, locus G6D	NM_021246	LY6G6D	51
calcium binding protein P22	NM_007236	CHP	52

CDC14 cell division cycle 14 homolog B (S. cerevisiae)	NM_003671 NM_033331	CDC14B	53 115
epiplakin 1	XM_372063	EPPK1	54
metallothionein 1X	NM_005952	MT1X	55
transforming growth factor, beta receptor II (70/80kDa)	NM_003242	TGFBR2	56
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 129
transmembrane 4 superfamily member 6	NM_003270	TM4SF6	58
pleckstrin homology domain containing, family B (evectins) member 1	NM_021200	PLEKHB1	59
apolipoprotein L, 1	NM_003661 NM_145343	APOL1	60 125
indoleamine-pyrrole 2,3 dioxygenase	NM_002164	INDO	61
forkhead box A2	NM_021784 NM_021784	FOXA2	62
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
mitochondrial solute carrier protein	NM_016612 NM_016612	MSCP	101
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

or from

Table 5

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 124
granzyme H (cathepsin G-like 2, protein h-CCPX)	NM_033423	GZMH	63
baculoviral IAP repeat-containing 3	NM_001165 NM_001165	BIRC3	64
Homo sapiens metallothionein 1H-like protein		AF333388 (Hs 382039)	135
KIAA0182 protein	NM_014615	KIAA0182	117
G protein-coupled receptor 56	NM_005682 NM_301524	GPR56	65 116

metallothionein 2A	NM_005953	MT2A	66
F-box only protein 21	NM_015002	FBXO21	67
erythrocyte membrane protein band 4.1-like 1	NM_012156	EPB41L1	68
hypothetical protein MGC21416	NM_173834	MGC21416	69
protein O-fucosyltransferase 1	NM_015352	POFUT1	70
metallothionein 1E (functional)	NM_175617	MT1E	71
troponin T1, skeletal, slow	NM_003283	TNNT1	72
chimerin (chimaerin) 2	NM_004067	CHN2	73
heterogeneous nuclear ribonucleoprotein H1 (H)	NM_005520	HNRPH1	74
ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle	NM_004046	ATP5A1	75
eukaryotic translation initiation factor 5A	NM_001970	EIF5A	76
perforin 1 (pore forming protein)	NM_005041	PRF1	77
OGT(O-Glc-NAc transferase)-interacting protein 106 KDa	NM_014965	OIP106	78
DEAD (Asp-Glu-Ala-Asp) box polypeptide 27	NM_017895	DDX27	79
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
hypothetical protein FLJ20232	NM_019008	FLJ20232	99
apolipoprotein L, 2	NM_030882, NM_145343	APOL2	100 120
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

or from

Table 6

Gene name	Ref seq	Gene sym-bol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
metallothionein 1G	NM_005950	MT1G	36
splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 129
vacuolar protein sorting 35 (yeast)	NM_018206	VPS35	80
tripartite motif-containing 44	NM_017583	TRIM44	81

transmembrane, prostate androgen induced RNA	NM_020182 NM_199169 NM_199170	TMEPAI	82 127 128
dynein, cytoplasmic, light polypeptide 2A	NM_014183 NM_177953	DNCL2A	83 122
leucine aminopeptidase 3	NM_015907	LAP3	84
chromosome 20 open reading frame 35	NM_018478 NM_033542	C20orf35	85 118
solute carrier family 38, member 1	NM_030674	SLC38A1	86
CGI-85 protein	NM_016028	CGI-85	87
death associated transcription factor 1	NM_022105, NM_080796	DATF1	88 121
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
sestrin 1	NM_014454	SESN1	90
hypothetical protein FLJ20315	NM_017763	FLJ20315	91
hypothetical protein FLJ20647	NM_017918	FLJ20647	92
membrane protein expressed in epithelial-like lung adenocarcinoma	NM_024792	CT120	93
DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide	NM_014314	RIG-I	94
keratin 23 (histone deacetylase inducible)	NM_015515	KRT23	95
UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6 (GalNAc-T6)	NM_007210	GALNT6	96
aryl hydrocarbon receptor nuclear translocator-like 2	NM_020183	ARNTL2	97
apobec-1 complementation factor	NM_014576 NM_138932	ACF	98 119
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

Another embodiment of the invention concerning the determination of microsatellite status is based on the expression pattern of at least 2 genes, such as at least 3 genes, such as at least 4 genes, such as at least 5 genes, such as at least 6 genes, such as at least 7 genes, such as at least 8 genes, such as at least 9 genes selected from the group of genes listed in Table 7 below.

Table 7

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 124
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

Another embodiment of the invention concerning the determination of microsatellite status is based on the expression pattern of at least 2 genes, such as at least 3 genes, such as at least 4 genes, such as at least 5 genes, such as at least 6 genes, such as at least 7 genes selected from the group of genes listed in Table 8 below.

Table 8

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
Splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 124
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
hypothetical protein FLJ20618	NM_017903	FLJ20618	102

One embodiment of the invention concerning the determination of microsatellite status is based on the expression pattern of at least one gene, for example two genes, for example 3 genes, such as 4 genes selected from the group of genes listed below in Table 9 and at least one gene, for example two genes, for example 3 genes, such as 4 genes, for example 5 genes selected from the group of genes listed below in Table 10

Table 9

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
Splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43

and

Table 10

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 124
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

Another embodiment relates to the determination of microsatellite status is based on the expression pattern of at least one gene, for example 2 genes, for example 3 genes, such as 4 genes selected from the group of genes that are down regulated in MSS colon cancers compared to MSI colon cancers listed below in Table 11, and at least one gene, for example 2 genes, for example 3 genes, such as 4 genes, for example 5 genes selected from the group of genes that are up-regulated in MSS colon cancers compared to MSI colon cancers listed below in Table 12.

Table 11

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
Splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43

Table 12

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 124
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

Sporadic or hereditary classification using gene expression patterns

One embodiment of the invention relates to a method of determining the hereditary or sporadic nature of cancer as the prognostic marker in an individual having contracted cancer based on determination of the expression pattern of at least 2 genes, such as at least 3 genes, such as at least 4 genes, such as at least 5 genes, such as at least 6 genes, such as at least 7 genes, such as at least 8 genes, such as at least 9 genes, such as at least 10 genes selected from the group of genes listed in Table 13.

Table 13

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
Homeo box C6	NM_004503	HOXC6	105
Piwi – like 1	NM_004764.2	PIWIL1	106
Mut L homolog 1	NM_00249.2	MLH1	107
Collapsin response mediator protein 1	NM_001313.2	CRMP1	108
Homeo box B2	NM_002145.2	HOXB2	109
Pyrroline-5-carboxylate synthetase (glutamate gamma-semialdehyd synthetase)	NM_002860.2	PYCS/ADH18A1	110
TGFB inducible early growth response	NM_005655.1	TIEG	111
Checkpoint with forkhead and ring finger domains	NM_018223.1	CHFR	112
Hypothetical protein FLJ13842	NM_024645.1	FLJ13842	113
Phosphoprotein regulated by mitogenic pathways	NM_025195.1	C8FW	114

In a further embodiment of the invention the determination of the hereditary or sporadic nature of cancer is based on the expression pattern of at least 2 genes as listed in Table 14.

Table 14

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
Piwi – like 1	NM_004764.2	PIWIL1	105
Mut L homolog 1	NM_00249.2	MLH1	106

In one specific embodiment the MLH1 gene is down regulated in sporadic disease. In yet another embodiment the PIWIL1 is expressed in lower amounts in hereditary cancer.

Gene names and definition

The genes according to the present invention are identified by their gene name, gene symbol and a reference sequence number (RefSeq). Furthemore, the genes listed have been assigned a SEQ ID No and the sequence is submitted in the accompanying sequence listing. The reference sequence number refers to the Reference Sequence collection prepared by the the National Center for Biotechnologic Information (NCBI) and where a comprehensive set of sequences for major research organisms is provided, see http://www.ncbi.nlm.nih.gov/RefSeq.

Sample preparation

A number of procedures for the isolation of nucleic acids (DNA, RNA, mRNA) from a sample are available and well known to a person skilled in the art. Genomic DNA may be isolated for detection of mutations of the genome, or for detection of copy number of a gene or a number of other applications which will be appreciated by the skilled artisan. RNA and especially mRNA will be isolated when expression levels of a gene or several genes are to be detected. The sample may be from fresh or frozen tissue as defined elsewhere herein.

Before analyzing the sample by for example oligonucleotide arrays or quantitative PCR one or more preparations of the sample may be performed. Often, sample preparations include extraction of intracellular material for example extraction of nucleic acids from whole cell samples, viruses and the like, amplification of nucleic acids, fragmentation, transcription, labelling and/or extension reactions. One or more of these preparation features may be incorporated readily into the present invention.

RNA extraction

Methods of isolating total RNA and/or mRNA are well known to those skilled in the art. In one embodiment total RNA is isolated from a given sample by extraction using acidic guanidinium-phenol-chloroform extraction. mRNA may be isolated form RNA or directly for example based on the polyadenylated tail of mRNA using oligo dT column chromatography or by using (dT)n magnetic beads (see, e.g. Sambrook et al., Molecular Cloning: A laboratory Manual 2nd Ed.), Vols 1-3, Cold Spring Harbour Laboratory (1989), or Current Protocols in Molecular Biology, F.Ausubel et al., ed. Greene Publishing and Wiley-Interscience, New York (1987).

PCR

PCR (Polymerase Chain Reaction) is a key technique in molecular genetics that permits the analysis of any short sequence of DNA or RNA without having to clone the short sequence first. PCR is used to reproduce (amplify) selected sections of DNA or RNA. PCR amplification generally involves the use of one strand of the target nucleic acid sequence as a template for producing a large number of complements to that sequence. Generally, two primer sequences complementary to different ends of a segment of the complementary strands of the target sequence hybridize with their respective strands of the target sequence, and in the presence of polymerase enzymes and nucleoside triphosphates, the primers are extended along the target sequence. The extensions are melted from the target sequence and the process is repeated, this time with the additional copies of the target sequence synthesized in the preceding steps. PCR amplification typically involves repeated cycles of denaturation, hybridization and extension reactions to produce sufficient amounts of the target nucleic acid. The first step of each cycle of the PCR involves the separation of the nucleic acid duplex formed by the primer extension. Once the strands are separated, the next step in PCR involves hybridizing the separated strands with primers that flank the target sequence. The primers are then extended to form complementary copies of the target strands. For successful PCR amplification, the primers are designed so that the position at which each primer hybridizes along a duplex sequence is such that an extension product synthesized from one primer, when separated from the template (complement), serves as a template for the extension of the other primer. The cycle of denaturation, hybridization, and extension is repeated as many times as necessary to obtain the desired amount of amplified nucleic acid.

In PCR methods, strand separation is normally achieved by heating the reaction to a sufficiently high temperature for a sufficient time to cause the denaturation of the duplex but not to cause an irreversible denaturation of the polymerase. Typical heat denaturation involves temperatures ranging from about 80°C to 105° C for times ranging from seconds to minutes. Strand separation, however, can be accomplished by any suitable denaturing method including physical, chemical, or enzymatic means. Strand separation may be induced by a helicase, for example, or an enzyme capable of exhibiting helicase activity.

In addition to PCR reactions, the methods and devices of the present invention are also applicable to a number of other reaction types, e.g., reverse transcription, nick translation, and the like.

PCR may also be used to quantify the amount of transcripts of a particular gene. Typically, quantitative PCR also called real-time PCR is performed on cDNA synthesised from mRNA. Methods of "quantitative" amplification are well known to those of skill in the art. For example, quantitative PCR involves simultaneously co-amplifying a known quantity of a control sequence using the same primers. This provides an internal standard that may be used to calibrate the PCR reaction. The high density array may then include probes specific to the internal standard for quantification of the amplified nucleic acid.

Thus, in one embodiment, this invention provides a method of detecting of the expression pattern formed by a number of genes. Generally, this method involves providing a high density array containing a multiplicity of probes of one or more particular length(s) that are complementary to subsequences of the mRNA transcribed by the target gene. In one embodiment the high density array may contain every probe of a particular length that is complementary to a particular mRNA. The probes of the high density array are then hybridized with their target nucleic acid alone and then hybridized with a high complexity, high concentration nucleic acid sample that does not contain the targets complementary to the probes. Thus, for example, where the target nucleic acid is an RNA, the probes are first hybridized with their target nucleic acid alone and then hybridized with RNA made from a cDNA library (e.g., reverse transcribed polyA.sup.+ mRNA) where the sense of the hybridized RNA is opposite that of the target nucleic acid (to insure that the

high complexity sample does not contain targets for the probes). Those probes that show a strong hybridization signal with their target and little or no cross-hybridization with the high complexity sample are preferred probes for use in the high density arrays of this invention.

The method of measuring the level of expression of a number of genes forming a pattern is, however, not limited to the methods described herein, but includes any quantitative measurement.

Fragmentation

In addition, amplified sequences may be subjected to other post amplification treatments. For example, in some cases, it may be desirable to fragment the sequence prior to hybridization with an oligonucleotide array, in order to provide segments which are more readily accessible to the probes, which avoid looping and/or hybridization to multiple probes. Fragmentation of the nucleic acids may generally be carried out by physical, chemical or enzymatic methods that are known in the art.

Hybridization

Following sample preparation, the sample can be subjected to one or more different analysis operations. A variety of analysis operations may generally be performed, including size based analysis using, e.g., microcapillary electrophoresis, and/or sequence based analysis using, e.g., hybridization to an oligonucleotide array.

In the latter case, the nucleic acid sample may be probed using an array of oligonucleotide probes. Oligonucleotide arrays generally include a substrate having a large number of positionally distinct oligonucleotide probes attached to the substrate. These arrays may be produced using mechanical or light directed synthesis methods which incorporate a combination of photolithographic methods and solid phase oligonucleotide synthesis methods.

Detection

In the present context high density expression arrays may be used to determine the presence and/or amounts of gene expression products in a sample. Likewise, low

density expression arrays may be used to determine the expression pattern in a sample.

While high density expression arrays can be used, other techniques are also contemplated. These include other techniques for assaying for specific mRNA species, including RT-PCR and Northern Blotting, as well as techniques for assaying for particular protein products, such as ELISA, Western blotting, and enzyme assays. Gene expression patterns according to the present invention are determined by measuring any gene product of a particular gene, including mRNA and protein. A pattern may be for two or more genes.

RNA or protein can be isolated and assayed from a test sample using any techniques known in the art. RNA or protein can for example be isolated from fresh or frozen biopsy, from formalin-fixed tissue.

Expression of genes may in general be detected by either detecting mRNA from the cells and/or detecting expression products, such as peptides and proteins.

mRNA detection

The detection of mRNA of the invention may be a tool for determining the developmental stage of a cell type which is defined by its pattern of expression of messenger RNA. For example, in particular stages of cells, high levels of ribosomal RNA are found whereas relatively low levels of other types of messenger RNAs may be found. Where a pattern is shown to be characteristic of a stage, a stage may be defined by that particular pattern of messenger RNA expression. The mRNA population is a good determinant of developmental stage, will be correlated with other structural features of the cell. In this manner, cells at specific developmental stages will be characterized by the intracellular environment, as well as the extracellular environment. The present invention also allows the combination of definitions based, in part, upon antigens and, in part, upon mRNA expression.

In one embodiment, the two may be combined in a single incubation step. A particular incubation condition may be found which is compatible with both hybridization recognition and non-hybridization recognition molecules. Thus, e.g., an incubation condition may be selected which allows both specificity of antibody

binding and specificity of nucleic acid hybridization. This allows simultaneous performance of both types of interactions on a single matrix. Again, where developmental mRNA patterns are correlated with structural features, or with probes which are able to hybridize to intracellular mRNA populations, a cell sorter may be used to sort specifically those cells having desired mRNA population patterns.

It is within the general scope of the present invention to provide methods for the detection of mRNA. Such methods often involve sample extraction, PCR amplification, nucleic acid fragmentation and labeling, extension reactions, transcription reactions and the like.

DNA extraction

DNA extraction may be relevant in case possible mutations in the genes are to be determined in addition to the determination of expression of the genes.

For those embodiments where whole cells, or other tissue samples are being analyzed, it will typically be necessary to extract the nucleic acids from the cells or viruses, prior to continuing with the various sample preparation operations. Accordingly, following sample collection, nucleic acids may be liberated from the collected cells, viral coat, etc., into a crude extract, followed by additional treatments to prepare the sample for subsequent operations, e.g., denaturation of contaminating (DNA binding) proteins, purification, filtration, desalting, and the like.

Liberation of nucleic acids from the sample cells, and denaturation of DNA binding proteins may generally be performed by physical or chemical methods. For example, chemical methods generally employ lysing agents to disrupt the cells and extract the nucleic acids from the cells, followed by treatment of the extract with chaotropic salts such as guanidinium isothiocyanate or urea to denature any contaminating and potentially interfering proteins.

Alternatively, physical methods may be used to extract the nucleic acids and denature DNA binding proteins, such as physical protrusions within microchannels or sharp edged particles piercing cell membranes and extract their contents. Combinations of such structures with piezoelectric elements for agitation can provide suitable shear forces for lysis.

More traditional methods of cell extraction may also be used, e.g., employing a channel with restricted cross-sectional dimension which causes cell lysis when the sample is passed through the channel with sufficient flow pressure. Alternatively, cell extraction and denaturing of contaminating proteins may be carried out by applying an alternating electrical current to the sample. More specifically, the sample of cells is flowed through a microtubular array while an alternating electric current is applied across the fluid flow. Subjecting cells to ultrasonic agitation or forcing cells through microgeometry apertures, thereby subjecting the cells to high shear stress resulting in rupture are also possible extraction methods.

Filtration

Following extraction, it will often be desirable to separate the nucleic acids from other elements of the crude extract, e.g., denatured proteins, cell membrane particles, salts, and the like. Removal of particulate matter is generally accomplished by filtration, flocculation or the like. Further, where chemical denaturing methods are used, it may be desirable to desalt the sample prior to proceeding to the next step. Desalting of the sample, and isolation of the nucleic acid may generally be carried out in a single step, e.g., by binding the nucleic acids to a solid phase and washing away the contaminating salts or performing gel filtration chromatography on the sample, passing salts through dialysis membranes, and the like. Suitable solid supports for nucleic acid binding include, e.g., diatomaceous earth, silica (i.e., glass wool), or the like. Suitable gel exclusion media, also well known in the art, may also be readily incorporated into the devices of the present invention, and is commercially available from, e.g., Pharmacia and Sigma Chemical.

Alternatively, desalting methods may generally take advantage of the high electrophoretic mobility and negative of DNA compared to other elements. Electrophoretic methods may also be utilized in the purification of nucleic acids from other cell contaminants and debris. Upon application of an appropriate electric field, the nucleic acids present in the sample will migrate toward the positive electrode and become trapped on the capture membrane. Sample impurities remaining free of the membrane are then washed away by applying an appropriate fluid flow. Upon reversal of the voltage, the nucleic acids are released from the membrane in a substantially purer form. Further, coarse filters may also be overlaid on the barriers

to avoid any fouling of the barriers by particulate matter, proteins or nucleic acids, thereby permitting repeated use.

Separation of contaminants by chromatography

In a similar aspect, the high electrophoretic mobility of nucleic acids with their negative charges, may be utilized to separate nucleic acids from contaminants by utilizing a short column of a gel or other appropriate matrix or gel which will slow or retard the flow of other contaminants while allowing the faster nucleic acids to pass.

This invention provides nucleic acid affinity matrices that bear a large number of different nucleic acid affinity ligands allowing the simultaneous selection and removal of a large number of preselected nucleic acids from the sample. Methods of producing such affinity matrices are also provided. In general the methods involve the steps of a) providing a nucleic acid amplification template array comprising a surface to which are attached at least 50 oligonucleotides having different nucleic acid sequences, and wherein each different oligonucleotide is localized in a predetermined region of said surface, the density of said oligonucleotides is greater than about 60 different oligonucleotides per 1 cm.sup.2, and all of said different oligonucleotides have an identical terminal 3' nucleic acid sequence and an identical terminal 5' nucleic acid sequence. b) amplifying said multiplicity of oligonucleotides to provide a pool of amplified nucleic acids; and c) attaching the pool of nucleic acids to a solid support.

For example, nucleic acid affinity chromatography is based on the tendency of complementary, single-stranded nucleic acids to form a double-stranded or duplex structure through complementary base pairing. A nucleic acid (either DNA or RNA) can easily be attached to a solid substrate (matrix) where it acts as an immobilized ligand that interacts with and forms duplexes with complementary nucleic acids present in a solution contacted to the immobilized ligand. Unbound components can be washed away from the bound complex to either provide a solution lacking the target molecules bound to the affinity column, or to provide the isolated target molecules themselves. The nucleic acids captured in a hybrid duplex can be separated and released from the affinity matrix by denaturation either through heat, adjustment of salt concentration, or the use of a destabilizing agent such as formamide, TWEEN.TM.-20 denaturing agent, or sodium dodecyl sulfate (SDS).

Affinity columns (matrices) are typically used either to isolate a single nucleic acid typically by providing a single species of affinity ligand. Alternatively, affinity columns bearing a single affinity ligand (e.g. oligo dt columns) have been used to isolate a multiplicity of nucleic acids where the nucleic acids all share a common sequence (e.g. a polyA).

Affinity matrices

The type of affinity matrix used depends on the purpose of the analysis. For example, where it is desired to analyze mRNA expression levels of particular genes in a complex nucleic acid sample (e.g., total mRNA) it is often desirable to eliminate nucleic acids produced by genes that are constitutively overexpressed and thereby tend to mask gene products expressed at characteristically lower levels. Thus, in one embodiment, the affinity matrix can be used to remove a number of preselected gene products (e.g., actin, GAPDH, etc.). This is accomplished by providing an affinity matrix bearing nucleic acid affinity ligands complementary to the gene products (e.g., mRNAs or nucleic acids derived therefrom) or to subsequences thereof. Hybridization of the nucleic acid sample to the affinity matrix will result in duplex formation between the affinity ligands and their target nucleic acids. Upon elution of the sample from the affinity matrix, the matrix will retain the duplexes nucleic acids leaving a sample depleted of the overexpressed target nucleic acids.

The affinity matrix can also be used to identify unknown mRNAs or cDNAs in a sample. Where the affinity matrix contains nucleic acids complementary to every known gene (e.g., in a cDNA library, DNA reverse transcribed from an mRNA, mRNA used directly or amplified, or polymerized from a DNA template) in a sample, capture of the known nucleic acids by the affinity matrix leaves a sample enriched for those nucleic acid sequences that are unknown. In effect, the affinity matrix is used to perform a subtractive hybridization to isolate unknown nucleic acid sequences. The remaining "unknown" sequences can then be purified and sequenced according to standard methods.

The affinity matrix can also be used to capture (isolate) and thereby purify unknown nucleic acid sequences. For example, an affinity matrix can be prepared that contains nucleic acid (affinity ligands) that are complementary to sequences not

previously identified, or not previously known to be expressed in a particular nucleic acid sample. The sample is then hybridized to the affinity matrix and those sequences that are retained on the affinity matrix are "unknown" nucleic acids. The retained nucleic acids can be eluted from the matrix (e.g. at increased temperature, increased destabilizing agent concentration, or decreased salt) and the nucleic acids can then be sequenced according to standard methods.

Similarly, the affinity matrix can be used to efficiently capture (isolate) a number of known nucleic acid sequences. Again, the matrix is prepared bearing nucleic acids complementary to those nucleic acids it is desired to isolate. The sample is contacted to the matrix under conditions where the complementary nucleic acid sequences hybridize to the affinity ligands in the matrix. The non-hybridized material is washed off the matrix leaving the desired sequences bound. The hybrid duplexes are then denatured providing a pool of the isolated nucleic acids. The different nucleic acids in the pool can be subsequently separated according to standard methods (e.g. gel electrophoresis).

As indicated above the affinity matrices can be used to selectively remove nucleic acids from virtually any sample containing nucleic acids (e.g., in a cDNA library, DNA reverse transcribed from an mRNA, mRNA used directly or amplified, or polymerized from a DNA template, and so forth). The nucleic acids adhering to the column can be removed by washing with a low salt concentration buffer, a buffer containing a destabilizing agent such as formamide, or by elevating the column temperature.

In one particularly preferred embodiment, the affinity matrix can be used in a method to enrich a sample for unknown RNA sequences (e.g. expressed sequence tags (ESTs)). The method involves first providing an affinity matrix bearing a library of oligonucleotide probes specific to known RNA (e.g., EST) sequences. Then, RNA from undifferentiated and/or unactivated cells and RNA from differentiated or activated or pathological (e.g., transformed) or otherwise having a different metabolic state are separately hybridized against the affinity matrices to provide two pools of RNAs lacking the known RNA sequences.

In a preferred embodiment, the affinity matrix is packed into a columnar casing. The sample is then applied to the affinity matrix (e.g. injected onto a column or applied to a column by a pump such as a sampling pump driven by an autosampler). The affinity matrix (e.g. affinity column) bearing the sample is subjected to conditions under which the nucleic acid probes comprising the affinity matrix hybridize specifically with complementary target nucleic acids. Such conditions are accomplished by maintaining appropriate pH, salt and temperature conditions to facilitate hybridization as discussed above.

For a number of applications, it may be desirable to extract and separate messenger RNA from cells, cellular debris, and other contaminants. As such, the device of the present invention may, in some cases, include an mRNA purification chamber or channel. In general, such purification takes advantage of the poly-A tails on mRNA. In particular and as noted above, poly- T oligonucleotides may be immobilized within a chamber or channel of the device to serve as affinity ligands for mRNA. Poly-T oligonucleotides may be immobilized upon a solid support incorporated within the chamber or channel, or alternatively, may be immobilized upon the surface(s) of the chamber or channel itself. Immobilization of oligonucleotides on the surface of the chambers or channels may be carried out by methods described herein including, e.g., oxidation and silanation of the surface followed by standard DMT synthesis of the oligonucleotides.

In operation, the lysed sample is introduced to a high salt solution to increase the ionic strength for hybridization, whereupon the mRNA will hybridize to the immobilized poly-T. The mRNA bound to the immobilized poly-T oligonucleotides is then washed free in a low ionic strength buffer. The poy-T oligonucleotides may be immobiliized upon poroussurfaces, e.g., porous silicon, zeolites silica xerogels, scintered particles, or other solid supports.

Light directed synthesis of oligonucleotide arrays

The basic strategy for light directed synthesis of oligonucleotide arrays is as follows. The surface of a solid support, modified with photosensitive protecting groups is illuminated through a photolithographic mask, yielding reactive hydroxyl groups in the illuminated regions. A selected nucleotide, typically in the form of a 3'-O-phosphoramidite-activated deoxynucleoside (protected at the 5' hydroxyl with a

photosensitive protecting group), is then presented to the surface and coupling occurs at the sites that were exposed to light. Following capping and oxidation, the substrate is rinsed and the surface is illuminated through a second mask, to expose additional hydroxyl groups for coupling. A second selected nucleotide (e.g., 5'-protected, 3'-O-phosphoramidite-activated deoxynucleoside) is presented to the surface. The selective deprotection and coupling cycles are repeated until the desired set of products is obtained. Since photolithography is used, the process can be readily miniaturized to generate high density arrays of oligonucleotide probes. Furthermore, the sequence of the oligonucleotides at each site is known. See, Pease, et al. Mechanical synthesis methods are similar to the light directed methods except involving mechanical direction of fluids for deprotection and addition in the synthesis steps.

For some embodiments, oligonucleotide arrays may be prepared having all possible probes of a given length. The hybridization pattern of the target sequence on the array may be used to reconstruct the target DNA sequence. Hybridization analysis of large numbers of probes can be used to sequence long stretches of DNA or provide an oligonucleotide array which is specific and complementary to a particular nucleic acid sequence. For example, in particularly preferred aspects, the oligonucleotide array will contain oligonucleotide probes which are complementary to specific target sequences, and individual or multiple mutations of these. Such arrays are particularly useful in the diagnosis of specific disorders which are characterized by the presence of a particular nucleic acid sequence.

Following sample collection and nucleic acid extraction, the nucleic acid portion of the sample is typically subjected to one or more preparative reactions. These preparative reactions include in vitro transcription, labeling, fragmentation, amplification and other reactions. Nucleic acid amplification increases the number of copies of the target nucleic acid sequence of interest. A variety of amplification methods are suitable for use in the methods and device of the present invention, including for example, the polymerase chain reaction method or (PCR), the ligase chain reaction (LCR), self sustained sequence replication (3SR), and nucleic acid based sequence amplification (NASBA).

The latter two amplification methods involve isothermal reactions based on isothermal transcription, which produce both single stranded RNA (ssRNA) and double stranded DNA (dsDNA) as the amplification products in a ratio of approximately 30 or 100 to 1, respectively. As a result, where these latter methods are employed, sequence analysis may be carried out using either type of substrate, i.e., complementary to either DNA or RNA.

Frequently, it is desirable to amplify the nucleic acid sample prior to hybridization. One of skill in the art will appreciate that whatever amplification method is used, if a quantitative result is desired, care must be taken to use a method that maintains or controls for the relative frequencies of the amplified nucleic acids.

Labelling of nucleic acids

The nucleic acids in a sample will generally be labeled to facilitate detection in subsequent steps. Labeling may be carried out during the amplification, in vitro transcription or nick translation processes. In particular, amplification, in vitro transcription or nick translation may incorporate a label into the amplified or transcribed sequence, either through the use of labeled primers or the incorporation of labeled dNTPs into the amplified sequence.

Hybridization between the sample nucleic acid and the oligonucleotide probes upon the array is then detected, using, e.g., epifluorescence confocal microscopy. Typically, sample is mixed during hybridization to enhance hybridization of nucleic acids in the sample to nucleoc acid probes on the array.

In some cases, hybridized oligonucleotides may be labeled following hybridization. For example, where biotin labeled dNTPs are used in, e.g., amplification or transcription, streptavidin linked reporter groups may be used to label hybridized complexes. Such operations are readily integratable into the systems of the present invention. Alternatively, the nucleic acids in the sample may be labeled following amplification. Post amplification labeling typically involves the covalent attachment of a particular detectable group upon the amplified sequences. Suitable labels or detectable groups include a variety of fluorescent or radioactive labeling groups well known in the art. These labels may also be coupled to the sequences using methods that are well known in the art.

Methods for detection depend upon the label selected. A fluorescent label is preferred because of its extreme sensitivity and simplicity. Standard labeling procedures are used to determine the positions where interactions between a sequence and a reagent take place. For example, if a target sequence is labeled and exposed to a matrix of different probes, only those locations where probes do interact with the target will exhibit any signal. Alternatively, other methods may be used to scan the matrix to determine where interaction takes place. Of course, the spectrum of interactions may be determined in a temporal manner by repeated scans of interactions which occur at each of a multiplicity of conditions. However, instead of testing each individual interaction separately, a multiplicity of sequence interactions may be simultaneously determined on a matrix.

Means of detecting labeled target (sample) nucleic acids hybridized to the probes of the high density array are known to those of skill in the art. Thus, for example, where a colorimetric label is used, simple visualization of the label is sufficient. Where a radioactive labeled probe is used, detection of the radiation (e.g with photographic film or a solid state detector) is sufficient.

In a preferred embodiment, however, the target nucleic acids are labeled with a fluorescent label and the localization of the label on the probe array is accomplished with fluorescent microscopy. The hybridized array is excited with a light source at the excitation wavelength of the particular fluorescent label and the resulting fluorescence at the emission wavelength is detected. In a particularly preferred embodiment, the excitation light source is a laser appropriate for the excitation of the fluorescent label.

The target polynucleotide may be labeled by any of a number of convenient detectable markers. A fluorescent label is preferred because it provides a very strong signal with low background. It is also optically detectable at high resolution and sensitivity through a quick scanning procedure. Other potential labeling moieties include, radioisotopes, chemiluminescent compounds, labeled binding proteins, heavy metal atoms, spectroscopic markers, magnetic labels, and linked enzymes. Another method for labeling may bypass any label of the target sequence. The target may be exposed to the probes, and a double strand hybrid is formed at those positions only. Addition of a double strand specific reagent will detect where

hybridization takes place. An intercalative dye such as ethidium bromide may be used as long as the probes themselves do not fold back on themselves to a significant extent forming hairpin loops. However, the length of the hairpin loops in short oligonucleotide probes would typically be insufficient to form a stable duplex.

Suitable chromogens will include molecules and compounds which absorb light in a distinctive range of wavelengths so that a color may be observed, or emit light when irradiated with radiation of a particular wave length or wave length range, e.g., fluorescers. Biliproteins, e.g., phycoerythrin, may also serve as labels.

A wide variety of suitable dyes are available, being primarily chosen to provide an intense color with minimal absorption by their surroundings. Illustrative dye types include quinoline dyes, triarylmethane dyes, acridine dyes, alizarine dyes, phthaleins, insect dyes, azo dyes, anthraquinoid dyes, cyanine dyes, phenazathionium dyes, and phenazoxonium dyes. A wide variety of fluorescers may be used either on their own or in conjunction with quencher molecules. Fluorescers of interest fall into a number of categories having certain primary functionalities. These primary functionalities include 1- and 2-aminonaphthalene, p,p'-diaminostilbenes, pyrenes, quaternary phenanthridine salts, 9-aminoacridines, p,p'-diaminobenzophenone imines, anthracenes, oxacarbocyanine, merocyanine, 3-aminoequilenin, perylene, bis-benzoxazole, bis-p-oxazolyl benzene, 1,2-benzophenazin, retinol, bis-3-aminopyridinium salts, hellebrigenin, tetracycline, sterophenol, benzimidzaolylphenylamine, 2-oxo-3-chromen, indole, xanthen, 7-hydroxycoumarin, phenoxazine, salicylate, strophanthidin, porphyrins, triarylmethanes and flavin. Individual fluorescent compounds which have functionalities for linking or which can be modified to incorporate such functionalities include, e.g., dansyl chloride; fluoresceins such as 3,6-dihydroxy-9-phenylxanthhydrol; rhodamineisothiocyanate; N-phenyl 1-amino-8-sulfonatonaphthalene; N-phenyl 2-amino-6-sulfonatonaphthalene; 4-acetamido-4-isothiocyanato-stilbene-2,2'-disulfonic acid; pyrene-3-sulfonic acid; 2-toluidinonaphthalene-6-sulfonate; N-phenyl, N-methyl 2-aminoaphthalene-6-sulfonate; ethidium bromide; stebrine; auromine-0,2-(9'-anthroyl)palmitate; dansyl phosphatidylethanolamine; N,N'-dioctadecyl oxacarbocyanine; N,N'-dihexyl oxacarbocyanine; merocyanine, 4-(3'pyrenyl)butyrate; d-3-aminodesoxy-equilenin; 12-(9'-anthroyl)stearate; 2-methylanthracene; 9-vinylanthracene; 2,2'-(vinylene-p-

phenylene)bisbenzoxazole; p-bis›2-(4-methyl-5-phenyl-oxazolyl)!benzene; 6-dimethylamino-1,2-benzophenazin; retinol; bis(3'-aminopyridinium) 1,10-decandiyl diiodide; sulfonaphthylhydrazone of hellibrienin; chlorotetracycline; N-(7-dimethylamino-4-methyl-2-oxo-3-chromenyl)maleimide; N-›p-(2-benzimidazolyl)-phenyl!maleimide; N-(4-fluoranthyl)maleimide; bis(homovanillic acid); resazarin; 4-chloro-7-nitro-2,1,3-benzooxadiazole; merocyanine 540; resorufin; rose bengal; and 2,4-diphenyl-3(2H)-furanone.

Desirably, fluorescers should absorb light above about 300 nm, preferably about 350 nm, and more preferably above about 400 nm, usually emitting at wavelengths greater than about 10 nm higher than the wavelength of the light absorbed. It should be noted that the absorption and emission characteristics of the bound dye may differ from the unbound dye. Therefore, when referring to the various wavelength ranges and characteristics of the dyes, it is intended to indicate the dyes as employed and not the dye which is unconjugated and characterized in an arbitrary solvent.

Fluorescers are generally preferred because by irradiating a fluorescer with light, one can obtain a plurality of emissions. Thus, a single label can provide for a plurality of measurable events.

Detectable signal may also be provided by chemiluminescent and bioluminescent sources. Chemiluminescent sources include a compound which becomes electronically excited by a chemical reaction and may then emit light which serves as the detectible signal or donates energy to a fluorescent acceptor. A diverse number of families of compounds have been found to provide chemiluminescence under a variety of conditions. One family of compounds is 2,3-dihydro-1,-4-phthalazinedione. The most popular compound is luminol, which is the 5-amino compound. Other members of the family include the 5-amino-6,7,8-trimethoxy- and the dimethylamino›ca!benz analog. These compounds can be made to luminesce with alkaline hydrogen peroxide or calcium hypochlorite and base. Another family of compounds is the 2,4,5-triphenylimidazoles, with lophine as the common name for the parent product. Chemiluminescent analogs include para-dimethylamino and -methoxy substituents. Chemiluminescence may also be obtained with oxalates, usually oxalyl active esters, e.g., p-nitrophenyl and a peroxide, e.g., hydrogen

peroxide, under basic conditions. Alternatively, luciferins may be used in conjunction with luciferase or lucigenins to provide bioluminescence.

Spin labels are provided by reporter molecules with an unpaired electron spin which can be detected by electron spin resonance (ESR) spectroscopy. Exemplary spin labels include organic free radicals, transitional metal complexes, particularly vanadium, copper, iron, and manganese, and the like. Exemplary spin labels include nitroxide free radicals.

Expression products

The present invention relates to the classification of cancer in a tissue sample, based on the expression pattern formed by expression products such as mRNA as described above. Furthermore, the invention also relates to determining expression products such as peptides and proteins. The expression products, peptides and proteins, may be detected by any suitable technique known to the person skilled in the art.

In a preferred embodiment the expression products are detected by means of specific antibodies directed to the various expression products, such as immunofluorescent and/or immunohistochemical staining of the tissue.

Immunohistochemical localization of expressed proteins may be carried out by immunostaining of tissue sections from the single tumors to determine which cells expressed the protein encoded by the transcript in question. The transcript levels were used to select a group of proteins supposed to show variation from sample to sample, making possible a rough correlation between level of protein detected and intensity of the transcript on the microarray.

For example tissue sections may be cut from paraffin-embedded tissue blocks, mounted, and deparaffinized by incubation at 80 C° for 10 min, followed by immersion in heated oil at 60 C for 10 min (Estisol 312, Estichem A/S, Denmark) and rehydration. Antigen retrieval is achieved in TEG (TrisEDTA-Glycerol) buffer using microwaves at 900 W. The tissue sections are cooled in the buffer for 15 min before a brief rinse in tap water. Endogenous peroxidase activity is blocked by incubating the sections with 1% H202 for 20 min, followed by three rinses in tap

water, 1 min each. The sections are subsequently soaked in PBS buffer for 2 min. The following steps are modified from the descriptions given by Oncogene Science Inc., in the Mouse Immunohistochemistry Detection System, XHCO1 (UniTect, Uniondale, NY, USA). Briefly, the tissue sections are incubated overnight at 4 C with primary antibody, followed by three rinses in PBS buffer for 5 min each. Afterwards, the sections are incubated with biotinylated secondary antibody for 30 min, rinsed three times with PBS buffer and subsequently incubated with ABC (avidin-biotinlylated horseradish peroxidase complex) for 30 min, followed by three rinses in PBS buffer. Staining is performed by incubation with AEC (3-amino-ethylcarbazole) for 10 min. The tissue sections are counter stained with Mayers hematoxylin, washed in tap water for 5 min. and mounted with glycerol-gelatin. Positive and negative controls may be included in each staining round with all antibodies.

In yet another embodiment the expression products may be detected by means of conventional enzyme assays, such as ELISA methods.

Furthermore, the expression products may be detected by means of peptide/protein chips capable of specifically binding the peptides and/or proteins assessed. Thereby an expression pattern may be obtained.

Levels of expression

In the present invention the pattern formed by the expression profiles of genes is used for classifying cancer. The presence and/or amount of a plurality of gene expression products are/is consequently determined. The level of expression of selected genes is compared between different cancer cells.

The expression level of a particular gene in one cell type may be determined to be at least one fold higher than the expression level in a second cell type. By a one fold higher expression level is meant that the level of expression is doubled, i.e. an expression level is determined to be of the value 2 in one cell type and thus a one fold higher expression level is 4. By the term two-fold is meant the value is tripled, i.e. an increase from 2 to 6.

Treatment

MSS occurs in 85% of the diagnosed colon cancers, whereas MSI occurs in 15 % of the diagnoses. The choice of treatment in relation to chemotherapeutic agents seems to be related to the microsatellite status of a given colon cancer. MSS cells are affected by treatment using fluorouracil-based drugs (Carethers JM, Hawn MT, Chauhan DP, et al., J Clin Invest 1996;98(1):199-206.; Carethers JM, Chauhan DP, Fink D, et al. Gastroenterology 1999;117(1):123-31; Koi M, Umar A, Chauhan DP, et al. Cancer Res 1994;54(16):4308-12., Hawn MT, Umar A, Carethers JM, et al. Cancer Res 1995;55(17):3721-5). MSI cells are hypersensitive to for example irinotecan and CPT (Hsiang YH, Lihou MG, Liu LF. Cancer Res 1989;49(18):5077-82; Jacob S, Aguado M, Fallik D, Praz F. Cancer Res 2001;61(17):6555-62). Thus, the ability to determine the microsatellite status of a colon cancer can facilitate the selection of a chemotherapeutic agent that will be effective in treatment of that particular colon cancer type.

One aspect of the present invention relates to treatment with anti cancer drugs following the determination of the microsatellite status of a cancer in a sample to be microsatellite stable (MSS) and a prognostic marker. Typically, the MSS status has been determined using the methods of the present invention. A preferred embodiment is treatment using fluorouracil-based drugs, such as 5-fluorouracil, N-methy-N'-nitro-N-nitrosoguanidine or 6-thioguanine. Yet another preferred embodiment is treatment with non-fluorouracil-based anti cancer drugs. Another preferred embodiment of the invention is treatment using anti cancer drugs used in any combination. However; the treatment may also be a combination of treatment using for example fluorouracil-based anti cancer drugs and/or non-fluorouracil-based anti cancer drugs in combination with other sorts of treatment such as surgical intervention, radiation therapy, radiofrequency ablation, immuno therapy, gene therapy.

Anti cancer drugs

In one embodiment of the method for treatment of an individual comprising the steps of selecting an individual having contracted colon cancer, wherein the microsatellite status is stable and treating the individual with anti cancer drugs that are suitable for the diagnosed nature of the cancer. Fluorouracil-based anti cancer drugs prevent cells from synthesizing DNA and RNA by interfering with the synthesis of nucleic

acids, thus disrupting the growth of the cancer cells. The drugs are therefore also called antimetabolites. The group of fluorouracil-based anti cancer drugs comprises a large number of drugs. One of the drugs currently used in the clinic is 5-fluorouracil that is used to treat several types of cancer including colon cancer. Another preferred drug of the same group is N-methy-N'-nitro-N-nitrosoguanidine, and in particular 6-thioguanine. The drugs belonging to the group of fluorouracil-based anti cancer drugs are not limited to the ones described above, but comprise all fluorouracil-based anti cancer drugs.

A preferred embodiment for the treatment of MSI cells is the use of non-fluorouracil-based drugs, for example Topoisomerase I-inhibitors such as irinotecan, such as CPT(CPT-II, Camptothecin).

Another embodiment regards the use of anti cancer drugs that are suitable for cancer treatment but are not fluorouracil-based drugs. The non-fluorouracil based group of drugs is a heterogenous group of drugs that comprises for example antibodies, chemotherapeutic drugs known as antineoplastic drugs, and antidote drugs such as folate. One preferred drug of the group is Cetuximab (Erbitux) which is used in the clinic to treat patients with advanced colorectal cancer that has spread to other parts of the body. Erbitux is a monoclonal antibody approved to treat this type of cancer. Another preferred embodiment is a combination treatment of Erbitux to be given intravenously with Irinotecan, another drug approved to fight colorectal cancer. Irinotecan and oxaliplatin are chemotherapy drugs that are given as a treatment for cancer in the colon or rectum and exemplify yet other preferred embodiments. Irinotecan and oxaliplatin belong to the group of antineoplastics. Leucovorin is another preferred drug belonging to the group of non-fluorouracil-based drugs, where leucovorin is the active form of the B complex vitamin, folate, and is used as an antidote to drugs that decrease levels of folic acid. Some treatments require what is called leucovorin rescue, because the drug used to treat the cancer has had an adverse effect on folic acid levels.

In a third embodiment is the sequential or combined use of drugs from either of the two groups of drugs.

Assay

A further aspect of the invention relates to an assay for classifying cancer in an individual having contracted cancer. At least one marker capable of detecting the microsatellite status in a sample is included in the assay together with at least one marker determining the prognostic marker. The determination of said microsatellite status marker and the determination of the prognostic marker may in the assay be determined sequentially or simultaneously.

In a preferred embodiment the assay comprises at least two markers, one for the microsatellite status and one for the prognostic marker.

The marker (s) is/are preferably specifically detecting a gene as identified herein, in particular the genes of the tables in the examples and as discussed above. However, the marker of microsatellite status may be determined by conventional microsatellite analysis as described elsewhere herein.

As discussed above the marker may be any nucleotide probe, such as a DNA, RNA, PNA, or LNA probe capable of hybridising to mRNA indicative of the expression level. The hybridisation conditions are preferably as described below for probes.

In another embodiment the marker is an antibody capable of specifically binding the expression product in question.

Detection of expression pattern

Patterns can be compared manually by a person or by a computer or other machine. An algorithm can be used to detect similarities and differences. The algorithm may score and compare, for example, the genes which are expressed and the genes which are not expressed. Alternatively, the algorithm may look for changes in intensity of expression of a particular gene and score changes in intensity between two samples. Similarities may be determined on the basis of genes which are expressed in both samples and genes which are not expressed in both samples or on the basis of genes whose intensity of expression are numerically similar.

Generally, the detection operation will be performed using a reader device external to the diagnostic device. However, it may be desirable in some cases, to incorporate the data gathering operation into the diagnostic device itself.

The detection apparatus may be a fluorescence detector, or a spectroscopic detector, or another detector.

Although hybridization is one type of specific interaction which is clearly useful for use in this mapping embodiment, antibody reagents may also be very useful.

Data gathering and analysis

Gathering data from the various analysis operations, e.g., oligonucleotide and/or microcapillary arrays, will typically be carried out using methods known in the art. For example, the arrays may be scanned using lasers to excite fluorescently labeled targets that have hybridized to regions of probe arrays mentioned above, which can then be imaged using charged coupled devices ("CCDs") for a wide field scanning of the array. Alternatively, another particularly useful method for gathering data from the arrays is through the use of laser confocal microscopy which combines the ease and speed of a readily automated process with high resolution detection.

Following the data gathering operation, the data will typically be reported to a data analysis operation. To facilitate the sample analysis operation, the data obtained by the reader from the device will typically be analyzed using a digital computer. Typically, the computer will be appropriately programmed for receipt and storage of the data from the device, as well as for analysis and reporting of the data gathered, i.e., interpreting fluorescence data to determine the sequence of hybridizing probes, normalization of background and single base mismatch hybridizations, ordering of sequence data in SBH applications, and the like.

It is an object of the present invention to provide a biological sample which may be classified or characterized by analyzing the pattern of specific interactions mentioned above. This may be applicable to a cell or tissue type, to the messenger RNA population expressed by a cell to the genetic content of a cell, or to virtually any sample which can be classified and/or identified by its combination of specific molecular properties.

Pharmaceutical composition

The invention also relates to a pharmaceutical composition for treating the classified cancer, such as colorectal tumors.

In one aspect the pharmaceutical composition comprises one or more of the peptides being expression products as defined above. In a preferred embodiment, the peptides are bound to carriers. The peptides may suitably be coupled to a polymer carrier, for example a protein carrier, such as BSA. Such formulations are well-known to the person skilled in the art.

The peptides may be suppressor peptides normally lost or decreased in tumor tissue administered in order to stabilise tumors towards a less malignant stage. In another embodiment the peptides are onco-peptides capable of eliciting an immune response towards the tumor cells.

In another aspect the pharmaceutical composition comprises genetic material, either genetic material for substitution therapy, or for suppressing therapy as discussed below.

In a third aspect the pharmaceutical composition comprises at least one antibody produced as described above.

In the present context the term pharmaceutical composition is used synonymously with the term medicament. The medicament of the invention comprises an effective amount of one or more of the compounds as defined above, or a composition as defined above in combination with pharmaceutically acceptable additives. Such medicament may suitably be formulated for oral, percutaneous, intramuscular, intravenous, intracranial, intrathecal, intracerebroventricular, intranasal or pulmonal administration. For most indications a localised or substantially localised application is preferred.

Strategies in formulation development of medicaments and compositions based on the compounds of the present invention generally correspond to formulation strategies for any other protein-based drug product. Potential problems and the guidance required to overcome these problems are dealt with in several textbooks, e.g.

"Therapeutic Peptides and Protein Formulation. Processing and Delivery Systems", Ed. A.K. Banga, Technomic Publishing AG, Basel, 1995.

Injectables are usually prepared either as liquid solutions or suspensions, solid forms suitable for solution in, or suspension in, liquid prior to injection. The preparation may also be emulsified. The active ingredient is often mixed with excipients which are pharmaceutically acceptable and compatible with the active ingredient. Suitable excipients are, for example, water, saline, dextrose, glycerol, ethanol or the like, and combinations thereof. In addition, if desired, the preparation may contain minor amounts of auxiliary substances such as wetting or emulsifying agents, pH buffering agents, or which enhance the effectiveness or transportation of the preparation.

Formulations of the compounds of the invention can be prepared by techniques known to the person skilled in the art. The formulations may contain pharmaceutically acceptable carriers and excipients including microspheres, liposomes, microcapsules, nanoparticles or the like.

The preparation may suitably be administered by injection, optionally at the site, where the active ingredient is to exert its effect. Additional formulations which are suitable for other modes of administration include suppositories, and, in some cases, oral formulations. For suppositories, traditional binders and carriers include polyalkylene glycols or triglycerides. Such suppositories may be formed from mixtures containing the active ingredient(s) in the range of from 0.5% to 10%, preferably 1-2%. Oral formulations include such normally employed excipients as, for example, pharmaceutical grades of mannitol, lactose, starch, magnesium stearate, sodium saccharine, cellulose, magnesium carbonate, and the like. These compositions take the form of solutions, suspensions, tablets, pills, capsules, sustained release formulations or powders and generally contain 10-95% of the active ingredient(s), preferably 25-70%.

The preparations are administered in a manner compatible with the dosage formulation, and in such amount as will be therapeutically effective. The quantity to be administered depends on the subject to be treated, including, e.g. the weight and age of the subject, the disease to be treated and the stage of disease. Suitable dosage

ranges are of the order of several hundred μg active ingredient per administration with a preferred range of from about 0.1 μg to 1000 μg, such as in the range of from about 1 μg to 300 μg, and especially in the range of from about 10 μg to 50 μg. Administration may be performed once or may be followed by subsequent administrations. The dosage will also depend on the route of administration and will vary with the age and weight of the subject to be treated. A preferred dosis would be in the interval 30 mg to 70 mg per 70 kg body weight.

Some of the compounds of the present invention are sufficiently active, but for some of the others, the effect will be enhanced if the preparation further comprises pharmaceutically acceptable additives and/or carriers. Such additives and carriers will be known in the art. In some cases, it will be advantageous to include a compound, which promotes delivery of the active substance to its target.

In many instances, it will be necessary to administrate the formulation multiple times. Administration may be a continuous infusion, such as intraventricular infusion or administration in more doses such as more times a day, daily, more times a week, weekly, etc.

Therapy

The invention further relates to a method of treating individuals suffering from a classified cancer, in particular for treating a colorectal tumor. In one embodiment of the present invention the tumor cell in question is a microsatellite stable (MSS) tumor cell. In a second embodiment the tumor cell in question is a microsatellite instable (MSI) tumor cell.

In one embodiment the invention relates to a method of substitution therapy, i.e. administration of genetic material generally expressed in MSI cells, but lost or decreased in classified MSS cancer cells. Thus, the invention relates to a method for reducing malignancy of a MSS tumor cell of the Dukes B class, said method comprising

obtaining at least one gene selected from genes being expressed at least one-fold higher in MSI cells than the amount expressed in said MSS tumor cell

introducing said at least one gene into the MSS tumor cell in a manner allowing expression of said gene(s)

The at least one gene is preferably selected individually from genes comprising a sequence as identified below

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43

In a preferred embodiment at least two different genes are introduced into the MSS tumor cell.

By the term one-fold is meant that the value is doubled, i.e. an expression level is 2 – a one-fold higher expression level is 4.

By the term two-fold is meant that the value is tripled, i.e. an expression level is 2 – a two-fold higher expression level is 6.

In another embodiment the invention relates to a method of substitution therapy, i.e. administration of genetic material generally expressed in MSS cells, but lost or decreased in classified MSI cancer cells. Thus, the invention relates to a method for reducing malignancy of a MSI tumor cell of the Dukes C class, said method comprising

obtaining at least one gene selected from genes being expressed at least one-fold higher in MSS cells than the amount expressed in said MSI tumor cell

introducing said at least one gene into the MSI tumor cell in a manner allowing expression of said gene(s).

The at least one gene is preferably selected individually from genes comprising a sequence as identified below

Gene name	Ref seq	Gene symbol	SEQ ID NO.:

protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 129
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

In a preferred embodiment at least two different genes are introduced into the MSI tumor cell.

The invention also relates to a method of treating individuals having contracted colon cancer, wherein the microsatellite status has been determined to be stable and the cancer has been determined to be hereditary in nature. Preferably, the determination of microsatellite status and the hereditary nature of the cancer has been determined according to the present invention, analyzing expression patterns as described herein. The method of treating individuals involves introducing at least one gene into the tumor cell, whereby the gene is being expressed. The at least one gene is selected individually from the group of genes listed below

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
Homo sapiens mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	NM_000251	MSH2	136
Mut L homolog 1	NM_00249.2	MLH1	107
Homo sapiens PMS1 postmeiotic segregation increased 1 (S. cerevisiae)	NM_000534	PMS1	137
Homo sapiens PMS2 postmeiotic segregation increased 2 (S. cerevisiae) (PMS2), mRNA	NM_000535	PMS2	138
Homo sapiens mutS homolog 6 (E. coli)	NM_000179	MSH6	139

In a preferred embodiment at least two different genes are introduced into the tumor cell.

The invention further relates to a method of treating individuals suffering from a classified cancer, in particular for treating a colorectal tumor. In one embodiment of the present invention the tumor cell in question is a microsatellite stable (MSS)

tumor cell. In a second embodiment the tumor cell in question is a microsatellite instable (MSI) tumor cell.

In yet another embodiment the invention relates to a method of administration of peptides to the tumor cell in question. The peptide is expressed by at least one gene selected from genes being expressed at least two-fold higher in tumor cells than in the amount expressed in the tumor cells in question. The tumor cell in question is MSI tumor cell or MSI tumor cell.

Thus, the invention relates to a method for reducing cell malignancy of a MSS tumor cell of the Dukes B class, said method comprising

contacting a a MSS tumor cell with at least one peptide expressed by at least one gene selected from genes being expressed in at least two-fold higher in MSI tumor cells than in the amount expressed in said MSS tumor cell.

The at least one peptide is selected individually from genes as listed below

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43

In a preferred embodiment the invention relates to a method for reducing cell malignancy of a MSI tumor cell of the Dukes C class, said method comprising

contacting a MSI tumor cell with at least one peptide expressed by at least one gene selected from genes being expressed in at least two-fold higher in MSS tumor cells than in the amount expressed in said MSI tumor cell.

The at least one peptide is selected individually from genes as listed below

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 129
hepatocellular carcinoma-associated antigen	NM_018487	HCA112	89

112			
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

In another aspect the invention relates to a therapy whereby genes generally correlated to disease are inhibited by one or more of the following methods:

One embodiment of the invention relates to a method for reducing malignancy of a MSS cell of the Dukes B class, said method comprising

obtaining at least one nucleotide probe capable of hybridising with at least one gene of a MSS tumor cell , said at least one gene being selected from genes being expressed in an amount at least one-fold lower in MSI tumor cells than the amount expressed in said MSS tumor cell, and

introducing said at least one nucleotide probe into the MSS tumor cell in a manner allowing the probe to hybridise to the at least one gene, thereby inhibiting expression of said at least one gene.

The probes are preferably selected from probes capable of hybridising to a nucleotide sequence comprising a sequence as identified below

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
protein kinase C binding protein 1	NM_012408 NM_183047	PRKCBP1	57 129
hepatocellular carcinoma-associated antigen 112	NM_018487	HCA112	89
hypothetical protein FLJ20618	NM_017903	FLJ20618	102
SET translocation (myeloid leukaemia-associated)	NM_003011.1	SET	103
ATPase, class II, type 9a	Xm_030577.9	ATP9a	104

Another embodiment of the invention relates to a method for reducing malignancy of a MSI tumor cell of the Dukes B class, said method comprising

obtaining at least one nucleotide probe capable of hybridising with at least one gene of a MSI tumor cell , said at least one gene being selected from genes being expressed in an amount at least one-fold lower in MSS tumor cells than the amount expressed in said MSI tumor cell, and

introducing said at least one nucleotide probe into the MSI tumor cell in a manner allowing the probe to hybridise to the at least one gene, thereby inhibiting expression of said at least one gene.

The probes are preferably selected from probes capable of hybridising to a nucleotide sequence comprising a sequence as identified below

Gene name	Ref seq	Gene symbol	SEQ ID NO.:
heterogeneous nuclear ribonucleoprotein L	NM_001533	HNRPL	11
metastais-associated gene family, member 2	NM_004739	MTA2	23
chemokine (C-X-C motif) ligand 10	NM_001565	CXCL10	35
splicing factor, arginine/serine-rich 6	NM_006275	SFRS6	43

These methods are preferably based on anti-sense technology, whereby the hybridisation of said probe to the gene leads to a down-regulation of said gene.

The down-regulation may of course also be based on a probe capable of hybridising to regulatory components of the genes in question, such as promoters.

In yet another embodiment the probes consists of the sequences identified above.

The hybridization may be tested in vitro at conditions corresponding to in vivo conditions. Typically, hybridization conditions are of low to moderate stringency. These conditions favour specific interactions between completely complementary sequences, but allow some non-specific interaction between less than perfectly matched sequences to occur as well. After hybridization, the nucleic acids can be "washed" under moderate or high conditions of stringency to dissociate duplexes that are bound together by some non-specific interaction (the nucleic acids that form these duplexes are thus not completely complementary).

As is known in the art, the optimal conditions for washing are determined empirically, often by gradually increasing the stringency. The parameters that can be changed to affect stringency include, primarily, temperature and salt concentration. In general, the lower the salt concentration and the higher the temperature, the higher the stringency. Washing can be initiated at a low temperature (for example, room temperature) using a solution containing a salt concentration that is equivalent to or lower than that of the hybridization solution. Subsequent washing can be carried out using progressively warmer solutions having the same salt concentration. As alternatives, the salt concentration can be lowered and the temperature maintained in the washing step, or the salt concentration can be lowered and the temperature increased. Additional parameters can also be altered. For example, use of a destabilizing agent, such as formamide, alters the stringency conditions.

In reactions where nucleic acids are hybridized, the conditions used to achieve a given level of stringency will vary. There is not one set of conditions, for example, that will allow duplexes to form between all nucleic acids that are 85% identical to one another; hybridization also depends on unique features of each nucleic acid. The length of the sequence, the composition of the sequence (for example, the content of purine-like nucleotides versus the content of pyrimidine-like nucleotides) and the type of nucleic acid (for example, DNA or RNA) affect hybridization. An additional consideration is whether one of the nucleic acids is immobilized (for example, on a filter).

An example of a progression from lower to higher stringency conditions is the following, where the salt content is given as the relative abundance of SSC (a salt solution containing sodium chloride and sodium citrate; 2X SSC is 10-fold more concentrated than 0.2X SSC). Nucleic acids are hybridized at 42°C in 2X SSC/0.1% SDS (sodium dodecylsulfate; a detergent) and then washed in 0.2X SSC/0.1% SDS at room temperature (for conditions of low stringency); 0.2X SSC/0.1% SDS at 42°C (for conditions of moderate stringency); and 0.1X SSC at 68°C (for conditions of high stringency). Washing can be carried out using only one of the conditions given, or each of the conditions can be used (for example, washing for 10-15 minutes each in the order listed above). Any or all of the washes can be repeated. As mentioned above, optimal conditions will vary and can be determined empirically.

In another aspect a method of reducing tumorigenicity relates to the use of antibodies against an expression product of a cell from the biological tissue. The antibodies may be produced by any suitable method, such as a method comprising the steps of

obtaining expression product(s) from at least one gene said gene being expressed as defined below,

immunising a mammal with said expression product(s) obtaining antibodies against the expression product.

Use

The methods described above may be used for producing an assay for classifying cancer in animal tissue.

Furthermore, the invention relates to the use of a peptide as defined above for preparation of a pharmaceutical composition for the treatment of a classified cancer in animal tissue.

Furthermore, the invention relates to the use of a gene as defined above for preparation of a pharmaceutical composition for the treatment of a classified cancer in animal tissue.

Also, the invention relates to the use of a probe as defined above for preparation of a pharmaceutical composition for the treatment of a biological condition in animal tissue.

Gene delivery therapy

The genetic material discussed above may be any of the described genes or functional parts thereof. The constructs may be introduced as a single DNA molecule encoding all of the genes, or different DNA molecules having one or more genes. The constructs may be introduced simultaneously or consecutively, each with the same or different markers.

The gene may be linked to the complex as such or protected by any suitable system normally used for transfection such as viral vectors or artificial viral envelope, liposomes or micellas, wherein the system is linked to the complex.

Numerous techniques for introducing DNA into eukaryotic cells are known to the skilled artisan. Often this is done by means of vectors, and often in the form of nucleic acid encapsidated by a (frequently virus-like) proteinaceous coat. Gene delivery systems may be applied to a wide range of clinical as well as experimental applications.

Vectors containing useful elements such as selectable and/or amplifiable markers, promoter/enhancer elements for expression in mammalian, particularly human, cells, and which may be used to prepare stocks of construct DNAs and for carrying out transfections are well known in the art. Many are commercially available.

Various techniques have been developed for modification of target tissue and cells in vivo. A number of virus vectors, discussed below, are known which allow transfection and random integration of the virus into the host. See, for example, Dubensky et al. (1984) Proc. Natl. Acad. Sci. USA 81:7529-7533; Kaneda et al., (1989) Science 243:375-378; Hiebert et al. (1989) Proc. Natl. Acad. Sci. USA 86:3594-3598; Hatzoglu et al., (1990) J. Biol. Chem. 265:17285-17293; Ferry et al. (1991) Proc. Natl. Acad. Sci. USA 88:8377-8381. Routes and modes of administering the vector include injection, e.g intravascularly or intramuscularly, inhalation, or other parenteral administration.

Advantages of adenovirus vectors for human gene therapy include the fact that recombination is rare, no human malignancies are known to be associated with such viruses, the adenovirus genome is double stranded DNA which can be manipulated to accept foreign genes of up to 7.5 kb in size, and live adenovirus is a safe human vaccine organisms.

Another vector which can express the DNA molecule of the present invention, and is useful in gene therapy, particularly in humans, is vaccinia virus, which can be rendered non-replicating (U.S. Pat. Nos. 5,225,336; 5,204,243; 5,155,020; 4,769,330).

Based on the concept of viral mimicry, artificial viral envelopes (AVE) are designed based on the structure and composition of a viral membrane, such as HIV-1 or RSV and used to deliver genes into cells in vitro and in vivo. See, for example, U.S. Pat. No. 5,252,348, Schreier H. et al., J. Mol. Recognit., 1995, 8:59-62; Schreier H et al., J. Biol. Chem., 1994, 269:9090-9098; Schreier, H., Pharm. Acta Helv. 1994, 68:145-159; Chander, R et al. Life Sci., 1992, 50:481-489, which references are hereby incorporated by reference in their entirety. The envelope is preferably produced in a two-step dialysis procedure where the "naked" envelope is formed initially, followed by unidirectional insertion of the viral surface glycoprotein of interest. This process and the physical characteristics of the resulting AVE are described in detail by Chander et al., (supra). Examples of AVE systems are (a) an AVE containing the HIV-1 surface glycoprotein gp160 (Chander et al., supra; Schreier et al., 1995, supra) or glycosyl phosphatidylinositol (GPI)-linked gp120 (Schreier et al., 1994, supra), respectively, and (b) an AVE containing the respiratory syncytial virus (RSV) attachment (G) and fusion (F) glycoproteins (Stecenko, A. A. et al., Pharm. Pharmacol. Lett. 1:127-129 (1992)). Thus, vesicles are constructed which mimic the natural membranes of enveloped viruses in their ability to bind to and deliver materials to cells bearing corresponding surface receptors.

AVEs are used to deliver genes both by intravenous injection and by instillation in the lungs. For example, AVEs are manufactured to mimic RSV, exhibiting the RSV F surface glycoprotein which provides selective entry into epithelial cells. F-AVE are loaded with a plasmid coding for the gene of interest, (or a reporter gene such as CAT not present in mammalian tissue).

The AVE system described herein in physically and chemically essentially identical to the natural virus yet is entirely "artificial", as it is constructed from phospholipids, cholesterol, and recombinant viral surface glycoproteins. Hence, there is no carry-over of viral genetic information and no danger of inadvertant viral infection. Construction of the AVEs in two independent steps allows for bulk production of the plain lipid envelopes which, in a separate second step, can then be marked with the desired viral glycoprotein, also allowing for the preparation of protein cocktail formulations if desired.

Another delivery vehicle for use in the present invention are based on the recent description of attenuated Shigella as a DNA delivery system (Sizemore, D. R. et al., Science 270:299-302 (1995), which reference is incorporated by reference in its entirety). This approach exploits the ability of Shigellae to enter epithelial cells and escape the phagocytic vacuole as a method for delivering the gene construct into the cytoplasm of the target cell. Invasion with as few as one to five bacteria can result in expression of the foreign plasmid DNA delivered by these bacteria.

A preferred type of mediator of nonviral transfection in vitro and in vivo is cationic (ammonium derivatized) lipids. These positively charged lipids form complexes with negatively charged DNA, resulting in DNA charged neutralization and compaction. The complexes endocytosed upon association with the cell membrane, and the DNA somehow escapes the endosome, gaining access to the cytoplasm. Cationic lipid:DNA complexes appear highly stable under normal conditions. Studies of the cationic lipid DOTAP suggest the complex dissociates when the inner layer of the cell membrane is destabilized and anionic lipids from the inner layer displace DNA from the cationic lipid. Several cationic lipids are available commercially. Two of these, DMRI and DC-cholesterol, have been used in human clinical trials. First generation cationic lipids are less efficient than viral vectors. For delivery to lung, any inflammatory responses accompanying the liposome administration are reduced by changing the delivery mode to aerosol administration which distributes the dose more evenly.

The following are non-limiting examples illustrating the present invention.

Examples

In the following two complete studies have been performed. The type of experiments performed in Study 1 is described in examples 1 to 6. Study 2 describes an earlier, additional study.

Study I

101 colorectal tumors were tested for microsatellite instability and their global gene transcription was measured using high-density oligonucleotide microarrays. Unsupervised and supervised classification methods were applied to visualize tumor classes and define sets of genes for classification. Real-time PCR was used to vali-

date the microarray data and to investigate platform independency on an independent set of 47 tumors.

Patients and biopsy specimens

A broad spectrum of patients representing different common groups of colorectal cancers were included (Table 15, Fig.1).

Table 15

Summary of clinicopathological and microsatellite features of colon cancer samples.

					Tumour stage			IHC negative stain	
Patient group (Danish,Finnish)	N	Median age (range)	Localization in colon right (Danish,Finnish)	left (Danish,Finnish)	N (Danish,Finnish) 0[a]	II	III	N (n tested) MLH1	MSH2
All cases	118 (44,75)	62.0 (32-87)	44 (7,37)	74 (36,38)	17 (6,11)	36 (14,22)	65 (23,42)	12 (56)	1 (56)
Sporadic microsatellite instable tumors[b]	20 (9,16)	66.8 (44-87)	15 (3,12)	9 (6,4)	-	8 (2,6)	12 (2,10)	6 (11)	0 (11)
Hereditary microsatellite instable tumors[b,c]	17 (4,13)	49.6 (32-75)	9 (2,7)	8 (2,6)	-	10 (2,8)	5 (1,4)	5 (7)	1 (7)
Microsatellite stable tumors	61 (30,36)	61.0 (36-85)	11 (0,11)	55 (30,25)	-	18 (10,8)	48 (20,28)	0 (38)	0 (38)

[a]normal biopsy taken from the resection edge of a tumor
[b]according to microsatellite analysis
[c]according to the Amsterdam criteria
[c]DK is Denmark, SF is Finland)

MSI and MSS tumors both from the right and left colon were used and the MSI tumors represent both sporadic and hereditary cases. Hereditary cases included both missense and truncating germline mutations. All specimens were verified by histology and contained more than 50% tumor tissue. The tumors were resected at 15 different clinics in Denmark and Finland. Patients with stage III disease received adjuvant chemotherapy in addition to curative surgery. None of the patients received pre-operative radiation or chemotherapy. Informed consent was obtained from patients to use their specimens and clinical and pathological data for research purposes and the local ethic committees approved the study.

Microsatellite-instability analysis

DNA was extracted from microdissected cancer tissue. Control DNA was extracted from blood samples when available, otherwise normal epithelium from the oral resection edge was used. Samples positive for markers BAT25 and BAT26 were scored as microsatellite instable whereas samples positive for only one of these markers were tested for further markers and scored as low-frequency MSI if none of these tested positive. Tumors with low-frequency MSI have similar clinical features as microsatellite stable tumors and were considered as such in this study. For determining the MSI status of the 47-tumors in the test set a pentaplex polymerase chain reaction with five quasimonomorphic mononucleotide repeats was used (uraweera N, Duval A, Reperant M, et al.. Gastroenterology 2002;123(6):1804-11).

RNA purification

Colon specimens were obtained fresh from surgery and were either immediately snap frozen in liquid nitrogen (Denmark) or transferred to –70°C freezers with as little delay as possible (Finland). The Finnish samples were stored as dry tissue, whereas the Danish samples were either embedded in OCD-compound or stored in a SDS/guadinium thiocyanate solution. Total RNA was isolated using Trizol (Invitrogen) or GenElute Kits (Sigma) following the manufacturers' instructions.

Gene expression analysis

Labeling of RNA, hybridization and scanning was performed as described elsewhere (Dyrskjot L, Thykjaer T, Kruhoffer M, et al. Nat Genet 2003;33(1):90-6) Biotin labeled cRNA was prepared from 10μg of total RNA and hybridized to the Human Genome U133A GeneChip (Affymetrix) containing 22.289 probesets. The readings from the quantitative scanning were analyzed by the Affymetrix Software MAS 5.0 and normalized using the quantile normalization procedure implemented in RMA (robust multiarray analysis (Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Nucleic Acids Res 2003;31(4):e15., Bolstad BM, Irizarry RA, Astrand M, Speed TP, Bioinformatics 2003;19(2):185-93)).

Statistical testing

The 101 samples were divided into four groups according to country of origin and MS status. To study if there was any systematic difference between samples from the two countries, a test statistic S_1/S_2 was used. S_1 is the sum of squared deviations when a gene effect, country effect and microsatellite effect is observed, and S_2 is

the same sum where a gene effect and microsatellite effect is present only. The significance is obtained by calculating S_1/S_2 of data achieved after one hundred random permutations of the country labels. The above test was also performed for each gene separately by considering the number of genes with a test value $S_1(g)/S_2(g)$ below a given threshold. The same method was used to evaluate differences between the MSI and MSS groups. These calculations were based on the expression level of 5.082 genes with a variance over all tumor samples larger than 0.2.

Microsatellite status classifier

The MSI classifier was based on the 5.082 genes defined above. A normal distribution was used with the mean dependent on the gene and the group. For each gene, the variation between the groups and the variation within the groups was calculated to select genes with a high ratio between these. To classify a sample, the sum over the genes of the squared distance from the sample value to the group mean was calculated, standardized by the variance, and assigned the sample to the nearest group. The sample to be classified was excluded when calculating group means and variances.

Real-time PCR

The procedures were performed as described (Birkenkamp-Demtroder K, Christensen LL, Olesen SH, et al. Cancer Res 2002;62(15):4352-63) except that short LNA (Locked Nucleic Acid) enhanced probes from a Human Probe Library (Exiqon™) was used. All samples were normalized to GAPDH as this gene has been reported to be constantly expressed in colorectal cancer samples (Andersen CL, Ledet JL, Orntoft TF, Cancer Res 2004;In press).

Classification of new independent test samples based on real-time PCR

To translate the microarray-defined classifier to a PCR-platform, the nine-classifier genes were analyzed by quantitative PCR on a subset of 18 of the 101 tumor samples. The average for each gene and group of the microarray data were multiplied with a constant so that the total average was equal to the average of the corresponding log (Loeb LA., Cancer Res 1991;51(12):3075-9) transformed PCR values. This translation can be made because the normalized PCR values are expected to be proportional to the normalized array values, and on a log scale this becomes an additive difference. The difference is gene specific and is therefore estimated for

each gene separately. Thus, the variation obtained from the microarray data, and used in the classifier, can be used directly on the PCR platform.

Example 1

Hierarchical clustering

The phylogenetic tree resulting from hierarchical clustering is shown with relevant clinical and laboratory data (Fig. 1). The cluster analysis based on 1239 genes with a variation across all samples larger than 0.5 reveals that the samples are mainly separated in accordance to the microsatellite status. On the upper trunk we find two clusters represented mainly by normal biopsies (14/20) and MSS tumors (21/25), respectively. The six tumors among the normal samples were labeled in four different batches, excluding a labeling batch effect.

The lower trunk is divided into a MSI cluster (30/36) and a second MSS cluster (37/37). The MSI cluster contains three morphologically normal tissue specimens. Notably, there is no sign of separation between sporadic and hereditary MSI samples and right sided and left sided tumors are interspersed among each other. The two MSS clusters are either dominated by Danish samples (19/25) or by Finnish samples (26/37), which is likely to be due to differences in sampling and preparation in the two countries.

Based on these observations, we performed a series of statistical tests to evaluate if the observed separation of tumors into MSS and MSI groups as well as into Danish and Finnish groups were significant. Our test value S_1/S_2 was 0.914 between Danish and Finnish and 0.908 between MSI and MSS groups, as compared to minimum values of 0.966 and 0.963, respectively, in 100 permutations, thus demonstrating a very clear separation between the groups. This clear distinction of the groups are caused by many genes even at very strict criteria, i.e. low test statistic $S_1(g)/S_2(g)$ values (See Table 16).

Table 16

Permutation test of groups

Pseudo group	S1/S2 from data	Smaller values in 100 permutations	Minimum in 100 permutations
DK-SF	0.9146	0	0.9660
MSI-MSS	0.9084	0	0.9631

Permutation test of genes

Pseudo group		$S_1(j)/S_2(j)$ < 0.6	< 0.7	< 0.8	< 0.9
DK-SF	number of genes	22	114	425	1536
	max in 100 permutations	0	1	4	132
MSI-MSS	number of genes	49	151	461	1600
	max in 100 permutations	0	0	13	251

Example 2

Construction of a classifier for microsatellite instability status

To construct the classifier the expression profiles from 101 stage II and III tumors were employed. A maximum likelihood classifier was built in order to select a low number of genes resulting in the best possible separation of the two groups. The performance of the classifier was tested using 1-5082 genes and found to be stable showing 2-5 errors when using 4 to several hundred genes (Fig. 2A). In the final classifier the 9 genes (Table 17) that were most frequently used in the crossvalidation were used which resulted in 3 errors (Fig. 2B). The stability of the classifier was tested by randomly chosen datasets (Fig.2C).

Table 17. Genes used for the classification of microsatellite status

Genechip Probe_ID	Gene name	Gene Symbol	Array signal* Microsatellite Stable tumors	Array signal* Microsatellite instable tumors
202072	heterogeneous nuclear ribonucleoprotein L	HNRPL	208 ±73	776 ±340
203444	metastasis-associated 1-like 1	MTA1L1	45 ±13	104 ±36
206108	splicing factor, arginine/serine-rich 6	SFRS6	74 ±56	478 ±242
204533	chemokine (C-X-C motif) ligand 10	CXCL10	111 ±80	315 ±535
212062	ATPase, class II, type 9a	ATP9a	588 ±222	208 ±114
218345	hepatocellular carcinoma-associated antigen 112	HCA112	1261 ±603	446 ±271
222444	hypothetical protein FLJ20618	FLJ20618	776 ±193	338 ±168
210047	SET translocation (myeloid leukemia-associated)	SET	1351 ±298	478 ±201
209048	protein kinase C binding protein 1	PRKCBP1	294 ±113	158 ±79

* Array signal are median signal intensity values ±standard deviation.

The gene MTA1L1 shown in figures and tables in the examples section is in the context of the present invention identical to MTA2 and has been assigned SEQ ID NO.: 23.

Example 3

Stability of classification

The mean error rate for MSS tumors was 1.62% and for MSI tumors 6.0% for microsatellite instable tumors. More than 50% of the errors were related to three tumors (marked with an asterisk in figure 2B) of which two were wrongly classified in all permutation and one in 94%. The remaining errors were mainly caused by four tumors (marked + in figure 2B) with error rates of 40-47%. In terms of sensitivity and specificity, the classification of MSS tumors could be made with a sensitivity of 98.4% and with a specificity of 94.0% (Table 18). After correction for prevalence of MSS tumors, the positive predictive value of MSS was calculated to be almost 98.5% and the negative predictive value to 91%.

Table 18

Permutation analysis of the microsatellite instability classifier

	Trainings set	**Test set**
	Errors in crossvalidation	Test errors
Microsatellite instable tumors	6.48% (n=25, range 0-4)	6.0% (n=9, range 0-3)
Microsatellite stable tumors	1.67% (n=30, range 0-3)	1.62% (n=37, range 0-4)
All	3.85% (n=55, range 0-5)	2.48% (n=46, range 0-4)

Example 4

Construction of a classifier for sporadic versus hereditary microsatellite instable tumors

In order to identify a gene set for identification of hereditary MSI tumors the 19 sporadic MSI samples and 18 HNPCC MSI samples in the data set were subjected to supervised classification as described above (Fig.3A, B). The mismatch repair gene *MLH1* show a general downregulation in sporadic disease whereas *PIWIL1* is lower expressed in hereditary cases (Fig. 3C). Using these two genes only one error occurred: a sporadic MSI tumor was classified as hereditary. Based on t-test 500 permutations were performed to test the significance of these two genes as marker genes and found both genes highly significant with estimated p-values < 0.005.

In addition to identifying MSI tumors, we were also able to classify these as being sporadic or inherited based on the expression of only two genes, *MLH1* and *PIWIL1*. The classification only showed one misclassification out of 37. This single case did not have any family history of CRC but the family is very small and the patient was diagnosed at the age of 32 and may thus represent a missed HNPCC case. A majority of sporadic and about half of hereditary microsatellite instable colorectal cancers are caused by inactivation of *MLH1*. In sporadic tumors this is mostly caused by biallelic promotor hypermethylation whereas somatic mutation or loss of heterozygosity of the wild-type allele is significant mechanisms in hereditary tumors. As a result, the *MLH1* expression level in sporadic disease in strongly compromised whereas one or two alleles of *MLH1* in HNPCC are transcribed although encoding a mutated protein. *PIWIL1* is a member of the human Argonaute family that contain a conserved RNA-binding PAZ domain and may be involved in the development and maintenance of stem cells through the RNA-mediated gene-quelling mechanisms associated with DICER (Yuan Z, Sotsky Kent T, Weber TK., Oncogene 2003;22(40):6304-10., Sasaki T, Shiohama A, Minoshima S, Shimizu N., Genomics 2003;82(3):323-30. The association of this gene with hereditary MSI tumors is unknown and needs further investigation.

Example 5

Cross platform classification

The gene expression level by was measured real-time PCR of the nine classifier genes in a subset of 18 MSS and MSI samples. Median centered and scaled PCR data gave the same overall picture as clustered array data from the 18 samples (Fig. 4A). As *SET* and *ATP9a* did not work well in the PCR reaction, and only seven of the nine classifier genes (*HNRPL*, *MTA1L1*, *SFR6*, *CXCL10*, *HCA112*, *FLJ20618* and *PRKCBP1*) were used and quantified the transcripts from these genes by real-time PCR in a new independent sample set of 47 tumors containing 35 MSS and 12 MSI. The classifier correctly classified 45 of 47 tumors with only two MSI tumors being classified as MSS (Fig. 4B).

Example 6

Relation between microsatellite-instability status, stage and survival

The 9-gene classifier was used to classify 36 patients with Stage II tumors and found that 17 were MSI and 19 MSS. The overall survival was highly significantly related to the classification since 10 of 11 patients that died within five years belonged to the MSS group (P=0.0014) (Fig. 5A). Thus, as expected the classifier clearly proved to be a strong predictor of survival in stage II disease and probably can be used to select those MSS patients who may benefit from adjuvant chemotherapy.

Among 65 patients with Stage III tumors receiving adjuvant chemotherapy, 16 were classified as MSI tumors and 49 as MSS tumors. As 6 MSI and 30 MSS patients died within five years of follow-up there was no significant difference in overall survival between these groups (P=0.55) (Fig. 5B). A trend was that the patients with MSI tumors showed a poorer short-term survival than those with MSS tumors, contrary to stage II patients. This difference may be in accordance with a recent large study which showed that chemotherapy only benefit the MSS tumor patients (Ribic CM, Sargent DJ, Moore MJ, et al., N Engl J Med 2003;349(3):247-57).

Study 2

Background

Colon cancers microsatellite instability status is a better marker for response to adjuvant chemotherapy with fluorouracil than tumour stage II and III. The majority of hereditary colorectal cancer cases are microsatellite instable. We investigated the possibility of classifying colon tumors based on gene expression in crude biopsies and correlated these to crude survival and investigated if the gene expression profile can also identify hereditary cases from sporadic cases.

Methods

Gene transcripts from tumour specimens were quantified using microarray technology. The tumors were clustered using unsupervised and supervised classification algorithms. Sets of genes were defined for classification of microsatellite instability status and sporadic verses hereditary microsatellite instable tumors. Real-time PCR was used to validate microarray data and to investigate platform dependency in a new independent set of 47 colorectal tumors.

Results

Unsupervised hierarchical clustering revealed that tumors were essentially separated according to microsatellite instability status. Supervised classification of the 97 tumor samples using a maximum likelihood classifier with a crossvalidation loop resulted in tree misclassification as compared to microsatellite analysis using from 106 genes and down to only seven genes. The stability of classification of colon tumors in relation to microsatellite status was tested by permutation analysis. The sensitivity for diagnosis of microsatellite stable tumors exceeded 99% with a specificity exceeding 96%. The positive and negative predictive values exceeded 95% and 98%, respectively. The classifier was demonstrated not to be platform dependent as it could successfully be reproduced by real-time PCR. This was further verified as the classifier also correctly classified 95.7% of a new independent set of 47 colorectal tumors using real-time PCR.

Based on microarray data we identified ten genes that were highly correlated with hereditary disease. Using down to two of these genes 36 of 37 microsatellite instable tumors could be correctly separated into sporadic and hereditary MSI-H colorectal tumors.

Crude survival according to microsatellite status as determined by the classifier, revealed that stage II colon receiving no adjuvant chemotherapy, that patient displaying microsatellite instability had significantly longer overall survival than patient exhibiting microsatellite stable tumors (P=0.0014).

By contrast, the patient with Dukes' C tumors displaying microsatellite instability did not have a significant increase in overall survival as compared to patient exhibiting microsatellite stable tumors (P=0.55).

Conclusion

Colon cancer can be stratified into two molecular distinct groups by quantification of the transcripts of 106 genes or even down to seven genes. The two groups are highly correlated with microsatellite stable (MSS) and microsatellite instable (MSI) tumors. The 7-gene classifier clearly proved to be a strong predictor of survival in Dukes B and it can be used to select patients who need adjuvant chemotherapy,

namely those classified as MSS. We demonstrate that this classification is also valid when performed by real-time PCR analysis allowing a fast diagnosis in a clinical setting. Finally, sporadic from hereditary cases in tumors exhibiting microsatellite instability can be identified based on gene expression monitoring.

Introduction

Colon is the fourth most frequently diagnosed malignancy and the second most common cause of cancer death in the western world. The standard treatment of colon cancer is advised according to tumor stage. Patient with Dukes' C colon cancer receives a flurouracil-based adjuvant systemic chemotherapy in addition to surgical resection of the tumor, whereas the treatment for Dukes' B patients is based alone on surgical resection.

There is accumulating evidence that these cancers belong to two distinct molecular types according to genetic alterations. The mutator phenotype featuring tumors with microsatellite instability (MSI) and the suppressor pathway displying chromosomal instability and microsatellite stable (MSS). MSI has been defined as a change of any length due to either insertions or deletions of repeating units in a microsatellite within a tumor compared to normal tissue and is caused by an underlying defect in the mismatch repair (MMR) system. (Boland et al, CR 1998, 58:5248). The MSI pathway may either be sporadic or hereditary (HNPCC) and whereas the disruption of the MMR system in sporadic MSI tumors is most often caused by somatic methylation of the MLH1 gene promoter more that 90% of HNPCC cancers are caused by germline mutations in MLH1 or MSH2.

The MSS pathway to cancer begins with the inactivation of tumor suppressor genes, such as APC/β-catenin genes, followed by activation of oncogenes and inactivation of additional tumor suppressor genes, commonly with a high frequency of allelic losses and cytogenetic abnormalities and abnormal DNA tumor content. Many studies have defined the pathoclinical trait of MSI and MSS tumors and found that MSI positive cancers are most frequently found in the right side of the colon, they tend to be of less differentiated, they tend to be larger in size, are often mucinous and often exhibit extensive infiltration by lymphocytes.

Crude survival data suggest that patients with HNPCC have a better prognosis than those with sporadic disease [48,49,50] and studies have also shown that MSI is an independent indicator of good prognosis [35,52,53]. Recently it was shown that MSS benefit from 5-FU treatment/leucovin treatment (New England J Med, august 2nd, 2003) whereas MSI cancer patients gained no advantage in survival.

Gene expression profiling has become an increasing used method for classification, outcome prediction, prediction of response (for a review see Dyrskjøt, expert opinion). Most such studies dealing with colon cancer have dealt with the identification of general tumor markers (Alon U; Levine AJ PNAS (1999); Kitahara O; Tsunoda T., Cancer Res (2001); Notterman DA; Levine AJ, Cancer Res (2001); Yanagava R; Nakamura (2001); Zou TT; Meltzer SJ, Oncogene (2002), Demtroeder CR (2002)), markers for benign adenomas versus adenocarcinomas (Lin YM; Nakamura Y, Oncogene (2002); Williams NS; Becerra C., Clin Cancer Res (2003)), staging (Frederiksen CM; Orntoft TF, J Cancer Res Clin Oncol (2003)), or liver metastasis (Takemasa I; Matsubara K., Biochem Biophys Res Commun (2001); Yanagawa R; Nakamura, Neoplasia (2001); Agrawal D; Yeatman T, J Natl Cancer Inst (2002)). One study has addressed the separation of low-frequency microsatellite instability tumors (MSI-L) from MSI and MSS (PCA) (Mori Y; Meltzer SJ, Cancer Res (2003)). The aim of this study was to build a general applicable and robust classifier based on gene expression to separate MSS from MSI tumors. To achieve such robustness the tumors for this study were collecting from 14-16 different clinics, RNA was isolated using different methods and labelled in several batches. Gene expression was measured by DNA microarrays of 101 Danish and Finnish tumors from primary colon cancer patients along with 17 normal biopsies.

Methods

Biological material From the Danish and Finnish CRC tissue banks 101 primary colon cancers and 17 macroscopically normal colon epithelium samples from the oral resection edge were chosen. Only adenocarcinomas from Dukes' stage B and C were included, however, these represented a broad spectrum of tumors in relation to location, heredity, microsatellite instability status, and origin of the patient. All tumors were collected in the period from 1994 to 2002, 68 tumor samples were collected at nine different clinics in Finland and 33 samples were collected at four different clinics in Denmark, 36 were Dukes' B, 67 Dukes' C, 41 were sporadic mi-

crosatellite highly instable (MSI-H) of which were 17 HNPCC, and 59 were sporadic microsatellite stable (MSS) (table 19). None of the patients received pre-operative radiation or chemotherapy.

Table 19

Summary of clinocopathological and microsatellite features of colon samples

Patient group		Median age	Localization in colon		Dukes' Stage n (DK,SF)			IHC negative stain N (n tested)	
n (DK,SF)		range	right (DK,SF)	left (DK,SF)	N[a]	B	C	MLH1	MSH2
All cases	119 (44,75)	62.0	45 (8,37)	74 (36,38)	17 (6,11)	36 (14,22)	66 (20,46)	12 (56)	1 (56)
MSI-H[b]	24 (9,16)	67.0	15 (3,12)	9 (6,4)	-	10 (3,7)	14 (5,9)	6 (11)	0 (11)
HNPCC[c]	17(4,13)	45.0	9 (2,7)	8 (2,6)	-	10 (2,8)	7 (2,5)	6 (8)	1 (8)
MSS	60 (25,35)	63.0	11 (0,11)	49 (25,24)	-	16 (9,7)	44 (16,28)	0 (37)	0 (37)

[a]normal biopsy taken from the resection edge of a tumor
[b]according to microsatellite analysis
[c]all tumors MSI-H[b]

Microsatellite-instability analysis. From all tumor samples available as paraffin blocks, ten sections were cut at 10μm and stained with haematoxylin. The first and last section was cut at 4 μm and stained with haematoxylin. These two sections were used for the identification of tumor and normal cells from each sample. Regions enriched in tumor cells (more than 90%) were microdissected from these sections and DNA was extracted using a Puregene DNA extraction kit (Gentra Systems, Minneapolis, MN). DNA from blood samples was used as control when available, otherwise normal tissue was microdissected from the tissue sections. The samples were analysed for microsatellite instability according to the NCI guidelines (Boland et al). Samples positive for markers BAT25 and BAT26 were scored as MSI-H. Samples positive for only one of these markers were tested for further markers and scored as MSI-L if none of these tested positive. Since MSI-L has similar clinical features as MSS these samples were considered as MSS in this study. In addition to microsatellite analysis all tumors from which paraffin blocks were available were tested for the presence of MLH1 and MSH2 protein by immunohistochemistry. None of the samples scored MSS were negative for either protein whereas six of the MSI scored samples were positive for both (Table 19).

RNA purification Colon specimens were obtained fresh from surgery and were immediately snap frozen in liquid nitrogen either as was, in OCD-compound or in a SDS/guadinium thiocyanate solution. Total RNA was isolated using RNAzol (WAK-Chemie Medical) or spin column technology (Sigma) following the manufactures' instructions.

Gene expression analysis These procedures were performed at described in detail elsewhere (Dyrskødt et al). Briefly, ten μg of total RNA was used as starting material for the target preparation as described. First and second strand cDNA synthesis was performed using the SuperScript II System (Invitrogen) according to the manufacturers' instructions except using an oligo-dT primer containing a T7 RNA polymerase promoter site. Labelled aRNA was prepared using the BioArray High Yield RNA Transcript Labelling Kit (Enzo) using Biotin labelled CTP and UTP (Enzo) in the reaction together with unlabeled NTP´s. Unincorporated nucleotides were removed using RNeasy columns (Qiagen). Fifteen μg of cRNA was fragmented, loading onto the Affymetrix HG_U133A probe array cartridge and hybridized for 16h. The arrays were washed and stained in the Affymetrix Fluidics Station and scanned using a confocal laser-scanning microscope (Hewlett Packard GeneArray Scanner G2500A). The readings from the quantitative scanning were analyzed by the Affymetrix Gene Expression Analysis Software (MAS 5.0) and normalized using RMA (robust multi array normalisation, Irizarry et al. 2002) in the statistical application R. Redundant probesets (as defined form Unigene build 168) with high correlation (>0.5) over all samples were removed, which reduced the dataset to approximately 14.400 probesets. This dataset was used a source for all further calculations in this manuscript.

Unsupervised agglomerative hierarchical clustering

For hierarchical cluster analysis 1239 genes with a variation across all samples greater than 0.5 were median-centred to a magnitude of 1. Samples and genes were then clustered using average linkage clustering with a modified Person correlation as similarity metric (Eisen et al., PNAS 95: 14863-14868, 1998). The cluster dendrogram was visualized with TreeView (Eisen).

Group testing

We make a statistical test where the p-value is evaluated through permutations. For each group and gene we calculate the average and the sum of squared deviations from the average. We then sum these over the genes and the groups:

This expression is calculated for joining DK with SF and MSI with MSS such that we end up with two groups. The sum of squared deviations is denoted S_2. As a test statistic we use S_1/S_2. A small value indicates that there is a real reduction in the deviations when going from 2 to 4 groups and thus the groups have a real significance. To judge if a value is significantly small we use permutations. For each of the four groups left when joining DK and SF we randomly allocate the members to a pseudo DK and pseudo SF in such a way that the number of members in each group are as in the original data.

$$S_1 = \sum_{groups} \sum_{genes} \left(X_{ij} - \bar{X}_{gr(i)j}\right)^2$$

To get an understanding of this separation we performed a test to see if this is caused by few genes or if many genes are involved. For this test we calculated $S_1 = \sum_{genes} S_1(gene)$ and similarly with $S_2 = \sum_{genes} S_2(gene)$. For each gene j we used the test statistic $S_1(j)/S_2(j)$ (Table 3).

Multidimentional scaling

We carried out multidimentional scaling on median-centered and normalized data using CMD—scale in the statistical application R and visualized in a two-dimentional plot.

Microsatellite status classifier

The readings from the quantitative scanning were analyzed by the Affymetrix Gene Expression Analysis Software (MAS 5.0) and normalized using RMA (robust multi array normalisation, Irizarry et al. 2002) in the statistical application R. Redundant probesets (as defined form Unigene build 168) with high correlation (>0.5) over all samples were removed, which reduced the dataset to approximately 14.400 probesets.

The microsatellite instability status classifier was based on a dataset of 4.266 genes. These genes result from the removal of genes with a variance over all tumor sam-

ples smaller than 0.2 and genes that separate Danish from Finnish samples with a t-value numerically greater than 2. We used a normal distribution with the mean dependent on the gene and the group (MSI, MSS). For each gene, we calculated the variation between the groups and the variation within the groups to select genes with a high ratio between these. To classify a sample, we calculated the sum over the genes of the squared distance from the sample value to the group mean, standardized by the variance and assigned the sample to the nearest group. The sample to be classified was excluded when calculating group means and variances.

Estimation of classifier stability

We validated the performance of the classifier by permutation. One hundred datasets consisting of 30 MSS samples and 25 MSI samples were randomly chosen by permutation for training of the classifier with the remaining samples in each case being assign to a testset. Averages over the 100 data sets of the number of errors in the cross-validation of the training set and in the test set were used as a measure of the precision of the classifier.

Real-time PCR (RT-PCR). The procedures were as described (Birkenkamp-Demtroder) except that we used short LNA (Locked Nucleic Acid) enhanced probes from a Human Probe Library (Exiqon™). In short, cDNA was synthesized from single samples some of which were previously analyzed on GeneChips. Reverse transcription was performed using Superscript II RT (Invitrogen). Real-time PCR analysis was performed on selected genes using the primers (DNA Technology) and probes (Exiqon, DK) described in figure legend X. All samples were normalized to GAPDH as described previously (Birkenkamp-Demtroder et. al. Cancer Res., *62*: 4352-4363, 2002).

Rebuilding of Classifier based on Real-Time PCR

The 79 tumors samples that were not analysed by real-time PCR were transformed into log ratios using one of the tumor samples as reference and used for training of the classifier. Then 23 samples of which 18 were also analyzed on arrays were equally transformed into log ratios using the same tumor sample as above as reference and tested. The idea behind this translation is that we expect the normalized PCR values to be proportional to the normalized array values, and on a log scale this becomes an additive difference. The difference is gene specific and is therefore

estimated for each gene separately. The variation obtained from the microarray data, and used in the classifier, can be used directly on the PCR platform.

Results

Hierarchical clustering

The clinical specimens used in this study were collected in two different countries from 14 different clinics in the period 1994 to 2001. The samples were selected to keep a balanced representation of microsatellite instable (MSI) and microsatellite stable (MSS) tumors from both the right- and left-sided colon. The MSI class was represented both by sporadic MSI and hereditary MSI (HNPCC) tumors. Only Dukes' B and Dukes' C tumor samples were included were selected (table 19). Before any attempt to divide a diverse sample collection into distinct classes analyzed the data for systematic bias that may have been introduces during the experimental procedures. A fast and easy way to discover both true distinct classes as well as systematic biases in the data is to perform a hierarchical clustering.

The phylogenetic tree resulting from hierarchical clustering on 1239 genes (Fig. 6) reveals that the main separating factor is microsatellite status. On the upper trunk we find two clusters represented mainly by normal biopsies (14/21) and MSS tumors (18/25), respectively. The lower trunk is divided into a MSI cluster (30/36) and a second MSS cluster (MSS2-cluster) (34/37). A closer inspection of the two MSS clusters unveil that one is dominated by Danish samples (19/25) and one by Finnish samples (26/37 check). Also, it is worth to notice that the MSI cluster contains a vast majority of Finnish samples (32/36) and that the sporadic MSI samples are interspersed among the hereditary samples. The normal biopsies cluster tight together with a slight tendency to separation according to origin. Tree normal samples cluster within the MSI cluster indicating that resection of these samples may have been to close to the tumor lesion.

Inspection of the gene cluster dendrogram shows that the two groups of MSS tumors are mainly separated by a large cluster of genes being upregulated in the Danish samples (data not shown) indicating that a systematic difference between Danish and Finnish samples.

Significance of observed groups

Based on these observations, we performed a series of test to evaluate if the observed separation of tumors into MSS and MSI as well as DK and SF are significant. For these tests the tumor samples were grouped into four virtual tumor-groups labelled, i.e. Danish MSI (MSI-DK), Danish MSS (MSS-DK), Finnish MSI (MSI-SF) and Finnish MSS (MSS-SF). Based on 5082 genes with a variance above 0.2, we tested if all four groups are significant or if some of the groups can be joined. We considered the two possibilities of joining DK and SF, and of joining MSI and MSS and made a statistical test where the p-value is evaluated through permutations. In 100 permutations of each group combination our test value S1/S2 is considerably smaller than in all permutation (Table 20) demonstrating a very clear separation between DK and SF and between MSI and MSS.

Table 20

Permutation test of groups

Pseudo group	S1/S2 from data	Smaller values in 100 permutations	Minimum in 100 permutations
DK-SF	0.9072795	0	0.962269
I-S	0.9166195	0	0.9583325

Such a clear distinction between groups may rely on a few highly separating genes or a general difference in the gene expression profile including many genes. For both the DK-SF and MSI-MSS the effect are caused by many genes even at very criteria, i.e. low test statistic $S_1(j)/S_2(j)$ values (Table 21).

Table 21

Permutation test of genes

		$S_1(j)/S_2(j)$			
Pseudo group		< 0.6	< 0.7	< 0.8	< 0.9
DK-SF	number of genes	36	136	522	1785
	max in 100 permutations	0	0	2	225
MSI-MSS	number of genes	17	103	399	1507
	max in 100 permutations	0	1	8	250

When a property is present that influences a large proportion of the genes this may obscure separation of clinical relevant features in unsupervised clustering. To visual-

ize the effect of such properties, we calculated distances by multidimensional scaling between samples with and without of 816 genes separating DK from SF with a t-value numerically greater than 2 (Fig 7). We see an improved separation of MSI and MSS with Danish and Finnish cases mixed. The MSI-DK samples are not completely separated as they are found both between the MSI-SF and the MSS samples. (These plots are not entirely unsupervised since the groups have been used to remove gene).

Construction of an MSI-MSS classifier

For the construction of a classifier we used the expression profiles from 97 tumors for which no ambiguity had been identified in relation to microsatellite status. The 816 genes separating DK from SF were excluded, as these would be unreliable for MS classification. We built a maximum likelihood classifier in order to select a minimum of genes giving the largest possible separation of the two groups. We tested the performance of the classifier using 1-1000 genes and found that it was stable showing 3-6 errors when using 4 – 400 genes. Of these 106 genes were especially suited for discrimination of MSS from MSI (table 22).

Table 22

AFFYID	*SYMBOL*	*LOCUS LINK*	*OMIM*	*REFSEQ*	*GENENAME*
1405_i_at	CCL5	6352	187011	NM_002985	chemokine (C-C motif) ligand 5
200628_s_at	WARS	7453	191050	NM_004184	tryptophanyl-tRNA synthetase
200814_at	PSME1	5720	600654	NM_006263	proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)
201641_at	BST2	684	600534	NM_004335	bone marrow stromal cell antigen 2
201649_at	UBE2L6	9246	603890	NM_004223	ubiquitin-conjugating enzyme E2L 6
201674_s_at	AKAP1	8165	602449	NM_003488	A kinase (PRKA) anchor protein 1
201762_s_at	PSME2	5721	602161	NM_002818	proteasome (prosome, macropain) activator subunit 2 (PA28 beta)
201884_at	CEACAM5	1048	114890	NM_004363	carcinoembryonic antigen-related cell adhesion molecule 5
201910_at	FARP1	10160	602654	NM_005766	FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)
201976_s_at	MYO10	4651	601481	NM_012334	myosin X
202072_at	HNRPL	3191	603083	NM_001533	heterogeneous nuclear ribonucleoprotein L
202203_s_at	AMFR	267	603243	NM_001144	autocrine motility factor receptor
202262_x_at	DDAH2	23564	604744	NM_013974	dimethylarginine dimethylaminohydrolase 2
202510_s_at	TNFAIP2	7127	603300	NM_006291	tumor necrosis factor, alpha-induced protein 2
202520_s_at	MLH1	4292	120436	NM_000249	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
202589_at	TYMS	7298	188350	NM_001071	thymidylate synthetase

202637_s_at	ICAM1	3383	147840	NM_000201	intercellular adhesion molecule 1 (CD54), human rhinovirus receptor
202678_at	GTF2A2	2958	600519	NM_004492	general transcription factor IIA, 2, 12kDa
202762_at	ROCK2	9475	604002	NM_004850	Rho-associated, coiled-coil containing protein kinase 2
203008_x_at	APACD	10190		NM_005783	ATP binding protein associated with cell differentiation
203315_at	NCK2	8440	604930	NM_003581	NCK adaptor protein 2
203335_at	PHYH	5264	602026	NM_006214	phytanoyl-CoA hydroxylase (Refsum disease)
203444_s_at	MTA2	9219	603947	NM_004739	metastais-associated gene family, member 2
203559_s_at	ABP1	26	104610	NM_001091	amiloride binding protein 1 (amine oxidase (copper-containing))
203773_x_at	BLVRA	644	109750	NM_000712	biliverdin reductase A
203896_s_at	PLCB4	5332	600810	NM_000933	phospholipase C, beta 4
203915_at	CXCL9	4283	601704	NM_002416	chemokine (C-X-C motif) ligand 9
204020_at	PURA	5813	600473	NM_005859	purine-rich element binding protein A
204044_at	QPRT	23475	606248	NM_014298	quinolinate phosphoribosyltransferase (nicotinate-nucleotide pyrophosphorylase (carboxylating))
204070_at	RARRES3	5920	605092	NM_004585	retinoic acid receptor responder (tazarotene induced) 3
204103_at	CCL4	6351	182284	NM_002984	chemokine (C-C motif) ligand 4
204131_s_at	FOXO3A	2309	602681	NM_001455	forkhead box O3A
204326_x_at	MT1X	4501	156359	NM_005952	metallothionein 1X
204415_at	G1P3	2537	147572	NM_002038, NM_022873	interferon, alpha-inducible protein (clone IFI-6-16)
204533_at	CXCL10	3627	147310	NM_001565	chemokine (C-X-C motif) ligand 10
204745_x_at	MT1G	4495	156353	NM_005950, NM_005950	metallothionein 1G
204780_s_at	TNFRSF6	355	134637	NM_000043, NM_152877, NM_152876, NM_152875, NM_152872, NM_152873, NM_152871	tumor necrosis factor receptor superfamily, member 6
204858_s_at	ECGF1	1890	131222	NM_001953	endothelial cell growth factor 1 (platelet-derived)
205241_at	SCO2	9997	604272	NM_005138	SCO cytochrome oxidase deficient homolog 2 (yeast)
205242_at	CXCL13	10563	605149	NM_006419	chemokine (C-X-C motif) ligand 13 (B-cell chemoattractant)
205495_s_at	GNLY	10578	188855	NM_006433, NM_006433	granulysin
205831_at	CD2	914	186990	NM_001767	CD2 antigen (p50), sheep red blood cell receptor
206108_s_at	SFRS6	6431	601944	NM_006275	splicing factor, arginine/serine-rich 6
206286_s_at	TDGF1	6997	187395	NM_003212	teratocarcinoma-derived growth factor 1
206461_x_at	MT1H	4496	156354	NM_005951	metallothionein 1H
206754_s_at	CYP2B6	1555	123930	NM_000767	cytochrome P450, family 2, subfamily B, polypeptide 6
206907_at	TNFSF9	8744	606182	NM_003811	tumor necrosis factor (ligand) superfamily, member 9

206918_s_at	RBM12	10137	607179	NM_006047, NM_006047	RNA binding motif protein 12
206976_s_at	HSPH1	10808		NM_006644	heat shock 105kDa/110kDa protein 1
207320_x_at	STAU	6780	601716	NM_004602, NM_004602, NM_017452, NM_017453	staufen, RNA binding protein (Drosophila)
207457_s_at	LY6G6D	58530	606038	NM_021246	lymphocyte antigen 6 complex, locus G6D
207993_s_at	CHP	11261	606988	NM_007236	calcium binding protein P22
208022_s_at	CDC14B	8555	603505	NM_003671, NM_003671, NM_033331	CDC14 cell division cycle 14 homolog B (S. cerevisiae)
208156_x_at	EPPK1	83481			epiplakin 1
208581_x_at	MT1X	4501	156359	NM_005952	metallothionein 1X
208944_at	TGFBR2	7048	190182	NM_003242	transforming growth factor, beta receptor II (70/80kDa)
209048_s_at	PRKCBP1	23613		NM_012408, NM_012408, NM_183047	protein kinase C binding protein 1
209108_at	TM4SF6	7105	300191	NM_003270	transmembrane 4 superfamily member 6
209504_s_at	PLEKHB1	58473	607651	NM_021200	pleckstrin homology domain containing, family B (evectins) member 1
209546_s_at	APOL1	8542	603743	NM_003661, NM_003661, NM_145343	apolipoprotein L, 1
210029_at	INDO	3620	147435	NM_002164	indoleamine-pyrrole 2,3 dioxygenase
210103_s_at	FOXA2	3170	600288	NM_021784, NM_021784	forkhead box A2
210321_at	GZMH	2999	116831	NM_033423	granzyme H (cathepsin G-like 2, protein h-CCPX)
210538_s_at	BIRC3	330	601721	NM_001165, NM_001165	baculoviral IAP repeat-containing 3
211456_x_at	AF333388				
212057_at	KIAA0182	23199		XM_050495	KIAA0182 protein
212070_at	GPR56	9289	604110	NM_005682	G protein-coupled receptor 56
212185_x_at	MT2A	4502	156360	NM_005953	metallothionein 2A
212229_s_at	FBXO21	23014		NM_015002, NM_015002	F-box only protein 21
212336_at	EPB41L1	2036	602879	NM_012156, NM_012156	erythrocyte membrane protein band 4.1-like 1
212341_at	MGC21416	286451		NM_173834	hypothetical protein MGC21416
212349_at	POFUT1	23509	607491	NM_015352, NM_015352	protein O-fucosyltransferase 1
212859_x_at	MT1E	4493	156351	NM_175617	metallothionein 1E (functional)
213201_s_at	TNNT1	7138	191041	NM_003283, NM_003283, XM_352926	troponin T1, skeletal, slow
213385_at	CHN2	1124	602857	NM_004067	chimerin (chimaerin) 2
213470_s_at	HNRPH1	3187	601035	NM_005520	heterogeneous nuclear ribonucleoprotein H1 (H)
213738_s_at	ATP5A1	498	164360	NM_004046	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle
213757_at	EIF5A	1984	600187	NM_001970	eukaryotic translation initiation factor 5A

214617_at	PRF1	5551	170280	NM_005041	perforin 1 (pore forming protein)
214924_s_at	OIP106	22906	608112	NM_014965	OGT(O-Glc-NAc transferase)-interacting protein 106 KDa
215693_x_at	DDX27	55661		NM_017895	DEAD (Asp-Glu-Ala-Asp) box polypeptide 27
215780_s_at	Hs.382039				
216336_x_at	AL031602				
217727_x_at	VPS35	55737	606931	NM_018206	vacuolar protein sorting 35 (yeast)
217759_at	TRIM44	54765		NM_017583	tripartite motif-containing 44
217875_s_at	TMEPAI	56937	606564	NM_020182, NM_020182, NM_199169, NM_199170	transmembrane, prostate androgen induced RNA
217917_s_at	DNCL2A	83658	607167	NM_014183, NM_014183, NM_177953	dynein, cytoplasmic, light polypeptide 2A
217933_s_at	LAP3	51056	170250	NM_015907	leucine aminopeptidase 3
218094_s_at	C20orf35	55861		NM_018478, NM_018478	chromosome 20 open reading frame 35
218237_s_at	SLC38A1	81539		NM_030674	solute carrier family 38, member 1
218242_s_at	CGI-85	51111		NM_016028, NM_016028	CGI-85 protein
218325_s_at	DATF1	11083	604140	NM_022105, NM_022105, NM_080796	death associated transcription factor 1
218345_at	HCA112	55365		NM_018487	hepatocellular carcinoma-associated antigen 112
218346_s_at	SESN1	27244	606103	NM_014454	sestrin 1
218704_at	FLJ20315	54894		NM_017763	hypothetical protein FLJ20315
218802_at	FLJ20647	55013		NM_017918	hypothetical protein FLJ20647
218898_at	CT120	79850		NM_024792	membrane protein expressed in epithelial-like lung adenocarcinoma
218943_s_at	RIG-I	23586		NM_014314	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide
218963_s_at	KRT23	25984	606194	NM_015515, NM_015515	keratin 23 (histone deacetylase inducible)
219956_at	GALNT6	11226	605148	NM_007210	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6 (GalNAc-T6)
220658_s_at	ARNTL2	56938		NM_020183	aryl hydrocarbon receptor nuclear translocator-like 2
220951_s_at	ACF	29974		NM_014576, NM_014576, NM_138932	apobec-1 complementation factor
221516_s_at	FLJ20232	54471		NM_019008	hypothetical protein FLJ20232
221653_x_at	APOL2	23780	607252	NM_030882, NM_030882	apolipoprotein L, 2
221920_s_at	MSCP	51312		NM_016612, NM_016612	mitochondrial solute carrier protein
222244_s_at	FLJ20618	55000		NM_017903	hypothetical protein FLJ20618

The minimum of three errors was found even using only 7 genes (Table 23).

Table 23. Genes used for the classification of MSS vs MSI tumors

Name	Symbol	Unigene	MSS	MSI
hepatocellular carcinoma-associated antigen 112	HCA112	Hs.12126	1261	653
metastasis-associated 1-like 1	MTA1L1	Hs.173043	45	91
chemokine (C-X-C motif) ligand 10	CXCL10	Hs.2248	104	274
heterogeneous nuclear ribonucleoprotein L	HNRPL	Hs.2730	194	630
hypothetical protein FLJ20618	FLJ20618	Hs.52184	776	388
splicing factor, arginine/serine-rich 6	SFRS6	Hs.6891	74	446
protein kinase C binding protein 1	PRKCBP1	Hs.75871	294	168

Classification of ambiguous samples

Application of the 7-gene classifier to the four samples showing ambiguity in the microsatellite analyses assigns all four to be microsatellite stable tumor class. Notably, all four showed expression levels of *Tumor Growth Factor β induced protein* (TFGBI), MLH1 and thymidylate synthase (TYMS) that are atypical for MSI tumors. Furthermore, these tumors were all from the left colon. Thus the misclassified tumors are clearly truly MSS or they belong to a yet undefined class of MSI tumors.

Stability of classification

To estimate the stability of the classifier based on all 97 tumor samples, we generated one hundred new classifiers based on randomly chosen datasets consisting of 30 MSS and 25 MSI samples. In each case the classifiers were tested with the remaining samples. The performance for each set was evaluated and averaged over all 100 training and test sets (Table 24). The mean error rate for MSS tumors was 0.52% and 1.38% for MSI tumors. The seven genes defined above were found to be those genes that were most frequently used in the crossvalidation loop. More than 50% of the errors were related to three tumors of which two were wrongly classified in all permutation and one in 94%. The remaining errors were mainly caused by four tumors with error rates of 40-47% showing that the former three samples are truly assigned contradictory to result from the microsatellite analysis and that four samples could not be assigned with confidence too any of the classes.

Table 24 Performance of the classifier

Trainings set	Test set
Errors in crossvalidation	Test errors

MSI	2.8% (n=25, range 0-6)	1.4% (n=10, range 0-4)
MSS	0.70% (n=30, range 0-3)	0.52% (n=29, range 0-2)
All	1.7% (n=55, range 1-7)	1.9% (n=39, range 0-5)

Table 25

Sensitivity, Specificity, and Predictive Value of Test for MSS based on the eight gene Classifier			
Positive for MSS	True = (0.9948*29)=28,8492		False = (0.138*10)= 1.38
Negative for MSS	False = (0.0052*29)= 0.1508		True = (0.962*10)= 9.62
Sensitivity	28.9507/29	=	99.5%
Specificity	9.62/10	=	96.2%
Positive predictive value	28.8492/30.2292	=	95.4%
Negative predictive value	9.62/9.7708	=	98.5%

*Based on a prevalence for MSS of 85%

Survival classifier

Using the same classification methods described above, we build classifiers for survival based on either all samples or the above defined groups of MSI-H and MSS. As seen in figure 10 a distinction of patient with good prognosis (>5 year survival) from patient with bad prognosis (< 5 years survival) can be achieved with higher precision and using only a fraction of the genes by first separating into MSI-H and MSS groups.

Construction of a classifier for sporadic versus hereditary microsatellite instable tumors

In order to identify a gene set for identification of hereditary microsatellite instable tumors we applied 19 sporadic microsatellite instable samples and 18 microsatellite instable samples to supervised classification as described above. We found ten genes we high scored for separation of sporadic MSI-H from hereditary MSI-H tumours (Table 26). In crossvalidation we found a minimum number of one error using two genes (Fig 9A) and were used in at least 36 of the 37 crossvalidation loops. The genes were: the mismatch repair gene MLH1 that show a general downregulation in sporadic disease and PIWIL1 that is lower expressed in hereditary cases (Fig 9B). Using these two genes only one error occurred: a sporadic microsatellite instable was classified as hereditary. Based on T-test we performed 500 permutations to test the significance of these two genes for marker genes and found both genes highly significant with p-values < 0.005.

Table 26

AFFYID	SYMBOL	LOCUSLINK	OMIM	REFSEQ	AFFYDESCRIPTION
206194_at	HOXC6	3223	142972	NM_004503	Homeo box C4
214868_at	PIWIL1	9271	605571	NM_004764.2	Piwi (Drosophila)-like 1
202520_s_at	MLH1	4292	120436	NM_000249.2	MutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2)
202517_at	CRMP1	1400	602462	NM_001313.2	Collapsin response mediator protein 1
205453_at	HOXB2	3212	142967	NM_002145.2	Homeo box B2 (HOXB2)
217791_s_at	PYCS/ADH18A1	5832	138250	NM_002860.2	Pyrroline-5-carboxylate synthetase (glutamate gamma-semialdehyde synthetase) (/PYCS/ADH18A1)
202393_s_at	TIEG	7071	601878	NM_005655.1	TGFB inducible early growth response (TIEG)
218803_at	CHFR	55743	605209	NM_018223.1	Checkpoint with forkhead and ring finger domains (CHFR)
219877_at	FLJ13842	79698		NM_024645.1	Hypothetical protein FLJ13842 (FLJ13842)
202241_at	C8FW	10221		NM_025195.2	Phosphoprotein regulated by mitogenic pathways (C8FW)

Cross platform classification

Real time PCR was applied both to verify the array data and examine if the 7-gene classifier would also perform on this platform. We chose 23 samples of which 18 were also analyzed on arrays. The correlation between the two platforms was high (data not shown). In order to test the performance of classification using PCR data we re-build our classifier with a 79 samples array dataset including only those tumors that were not analyzed with PCR. Two samples were classified in discordance with the microsatellite instability test of which one of them was ambiguously classified by the 7-gene array classifier.

Relation between microsatellite-instability status, stage and survival

Based on the 7-gene classifier, classification of 36 patients with Dukes' B tumors receiving no adjuvant chemotherapy, 18 were classified as MSI tumors and 18 as MSS tumors. The overall survival was highly significantly related to the classification since all nine patients that died within five years of follow-up were belonged to the

MSS group (P=0.0014) (Fig. 10A). Thus, the 7-gene classifier clearly proved to be a strong predictor of survival in Dukes B and it can be used to select patients who need adjuvant chemotherapy, namely those classified as MSS.

Among 65 patients with Dukes' C tumors receiving adjuvant chemotherapy, 17 were classified as MSI tumors and as 48 MSS tumors. Of these, 6 MSI and 27 MSS patients died within five years of follow-up meaning no significant difference in overall survival between these groups (P=0.55) (Fig. 10B). A trend was that the MSI showed a poorer short-term survival than the MSS, contrary to Dukes B patients. This difference can be attributed to the fact that a recent large study has shown that chemotherapy only benefit the MSS tumor patients, thus improving their survival to a level comparable to that which is characteristic of MSI tumor patients.

Clinical application of the discovery

In the clinic the 106 or less genes described can be used for predicting outcome of colorectal cancer when examined at the RNA level and also on the protein level as each gene identified is the project is transcribed to RNA that is further translated into protein. The genes can also be used determine which patient should be treated with chemotherapy as only non-microsatellite instable tumors will respond to 5-FU based therapy. Building classifiers can achieve a further stratification of patient with god and bad prognosis after stratification into microsatellite instable and stable tumors. The genes used to identify hereditary disease can be used to decide which patient should enter into sequencing analysis of mismatch repair genes.

The RNA determination can be made in any form using any method that will quantify RNA. The proteins can be measured with any method quantification method that can determine the level of proteins.

References

Agrawal D, Chen T, Irby R, Quackenbush J, Chambers AF, Szabo M, Cantor A, Coppola D, Yeatman TJ. Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. J Natl Cancer Inst. 2002 Apr 3;94(7):513-21.

Birkenkamp-Demtroder K, Christensen LL, Olesen SH, Frederiksen CM, Laiho P, Aaltonen LA, Laurberg S, Sorensen FB, Hagemann R, ORntoft TF. Gene expression in colorectal cancer. Cancer Res. 2002 Aug 1;62(15):4352-63.

Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res. 1998 Nov 15;58(22):5248-57. Review.

Chapusot C, Martin L, Bouvier AM, Bonithon-Kopp C, Ecarnot-Laubriet A, Rageot D, Ponnelle T, Laurent Puig P, Faivre J, Piard F. Microsatellite instability and intratumoural heterogeneity in 100 right-sided sporadic colon carcinomas. Br J Cancer. 2002 Aug 12;87(4):400-4.

Dyrskjot L, Thykjaer T, Kruhoffer M, Jensen JL, Marcussen N, Hamilton-Dutoit S, Wolf H, Orntoft TF. Identifying distinct classes of bladder carcinoma using microarrays. Nat Genet. 2003 Jan;33(1):90-6.

Frederiksen CM, Knudsen S, Laurberg S, Orntoft TF. Classification of Dukes' B and C colorectal cancers using expression arrays. J Cancer Res Clin Oncol. 2003 May;129(5):263-71.

Huang J, Qi R, Quackenbush J, Dauway E, Lazaridis E, Yeatman T. Effects of ischemia on gene expression.
J Surg Res. 2001 Aug;99(2):222-7.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003 Feb 15;31(4):e15.

Loukola A, Eklin K, Laiho P, Salovaara R, Kristo P, Jarvinen H, Mecklin JP, Launonen V, Aaltonen LA. Microsatellite marker analysis in screening for hereditary nonpolyposis colorectal cancer (HNPCC). Cancer Res. 2001 Jun 1;61(11):4545-9.

Markowitz S, Hines JD, Lutterbaugh J, Myeroff L, Mackay W, Gordon N, Rustum Y, Luna E, Kleinerman J.
Mutant K-ras oncogenes in colon cancers Do not predict Patient's chemotherapy response or survival. Clin Cancer Res. 1995 Apr;1(4):441-5.

Mori Y, Selaru FM, Sato F, Yin J, Simms LA, Xu Y, Olaru A, Deacu E, Wang S, Taylor JM, Young J, Leggett B, Jass JR, Abraham JM, Shibata D, Meltzer SJ. The impact of microsatellite instability on the molecular phenotype of colorectal tumors. Cancer Res. 2003 Aug 1;63(15):4577-82.

Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, Goldberg RM, Hamilton SR, Laurent-Puig P, Gryfe R, Shepherd LE, Tu D, Redston M, Gallinger S. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer.
N Engl J Med. 2003 Jul 17;349(3):247-57.